AD________________

Award Number: DAMD17-96-1-6012

TITLE: Computer-aided Classification of Malignant and Benign Lesions on Mammograms

PRINCIPAL INVESTIGATOR: Berkman Sahiner, Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, Michigan 48103-1274

REPORT DATE: May 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE May 2000	3. REPORT TYPE AND DATES COVERED Annual(1 May 99 - 30 Apr 00)

4. TITLE AND SUBTITLE
Computer-aided Classification of Malignant and Benign Lesions on Mammograms

5. FUNDING NUMBERS
DAMD17-96-1-6012

6. AUTHOR(S)
Berkman Sahiner, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, Michigan 48103-1274

E-MAIL:
berki@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Computerized classification methods were developed for characterization of mammographic lesions. For mass characterization, features related to the degree of spiculation were developed. These features were combined with morphological features related to the computer-segmented mass shape for classification using stepwise feature selection and linear discriminant analysis. For different views of the same mass, the malignancy score provided by the classifier were combined by averaging. The classification accuracy was measured using the area A_z under the receiver operating characteristics (ROC) curve. The trained classifier achieved a test A_z value of 0.87 on an independent data set of 45 masses. For microcalcification characterization, morphological features were extracted from computer-identified leisons. Morphological and texture features were combined using stepwise feature selection and linear discriminant analysis. The classifier was tested using the leave-one-case-out method. On a data set of 112 pairs of mammograms, the test A_z value of the computer was 0.83. In an ROC study, 7 experienced breast radiologists read the same 112 pairs of mammograms. The area A_z under the average ROC curve for radiologists was 0.71. The A_z value of the computer was higher than that of all radiologists, and the difference was statistically significant for three of the radiologists (p=0.03).

14. SUBJECT TERMS
Breast Cancer, Computer-aided diagnosis

15. NUMBER OF PAGES
79

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

June 22, 03

PI - Signature Date

(4) Table of Contents

(5) Introduction

Treatment of the breast cancer at an early stage is the most significant means of improving the survival rate of the patients. Mammography is currently the most sensitive method for detecting early breast cancer, and it is also the most practical for screening. However, the positive predictive value of mammographic diagnosis is only about 15%-30%. As the number of patients who undergo mammography increases, it will be increasingly important to improve the positive predictive value of mammography in order to reduce costs and patient discomfort. In this proposal, our goal is to investigate the problem of classifying mammographic lesions as malignant or benign using computer vision, automatic feature extraction, statistical classification, and artificial intelligence techniques. Our efforts are concentrated on the computer-aided classification of two kinds of breast abnormalities, masses and microcalcifications, which are the primary mammographic signs of malignancy. We are investigating computerized extraction of useful features for the differentiation of malignant and benign cases for both abnormalities, and the application of classical statistical classifiers and newly developed paradigms such as neural networks and genetic algorithms for the classification task. Our purposes are to i) improve existing techniques, devise new methods, and identify the preferred approaches for the classification of mammographic lesions, ii) show that computerized classification of mammographic lesions is feasible, and iii) develop a computerized program that can subsequently be shown to improve radiologists' classification of mammographic abnormalities.

(6) Body

In the fourth year (5/1/99-4/30/00) of this grant, we have performed the following studies:

(A) Development of spiculation features for classification of masses

Spiculations are important indicators of malignancy on mammograms. In the third year of the project, we had reported on the development of a spiculation detection method. After the spiculations were detected and segmented, the shape of the segmented mass was modified by appending the segmented spiculations to the core of the mass. Subsequently, morphological features were extracted from the segmented mass shape.

In the fourth year of the project, we extracted features directly related to the degree of spiculation of the mass. The extraction of these features is described next. Let (i_c,j_c) be a pixel on the mass contour. (Fig. 1). We first define a search region S as shown in Fig. 2. For each pixel (i,j) in S, we compute the angular difference θ between the image gradient direction at image pixel (i,j), and the direction of the vector joining pixels (i_c,j_c), and (i,j) (Fig. 1). If the pixel (i_c,j_c) lies on the path of a spiculation, then θ will be close to $\pi/2$ whenever the image pixel (i,j) is on the spiculation. Therefore, the distribution of θ, obtained from all image pixels (i,j) within the search region S will have a peak around $\pi/2$. If there is no spiculation, and if the gray levels in S are randomly distributed, then this distribution will be uniform. This was the basic idea behind the spiculation detection method reported last year. In the fourth year of the project, we computed the average of θ within S for all pixels on the mass boundary. In addition, the mass boundary was enlarged one pixel at a time, and this computation was repeated in a 30-pixel wide (3cm) ring around the segmented mass. A new image, called the spiculation likelihood map, was generated, in which the gray-level value for pixel (i_c,j_c) was the average of θ within the window S for pixel (i_c,j_c).

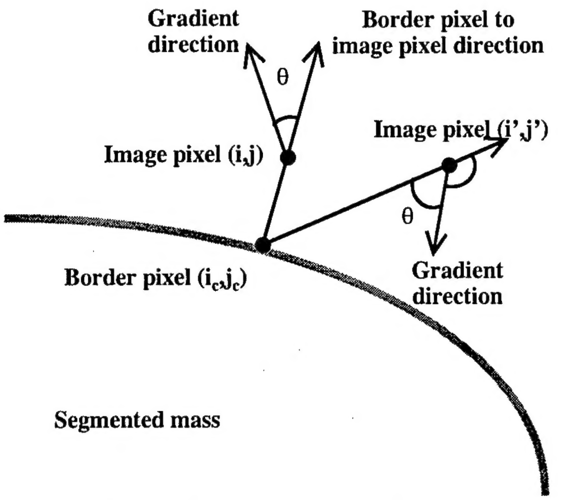


Figure 1. The definition of the pixel (i_c,j_c) and the angular difference θ.

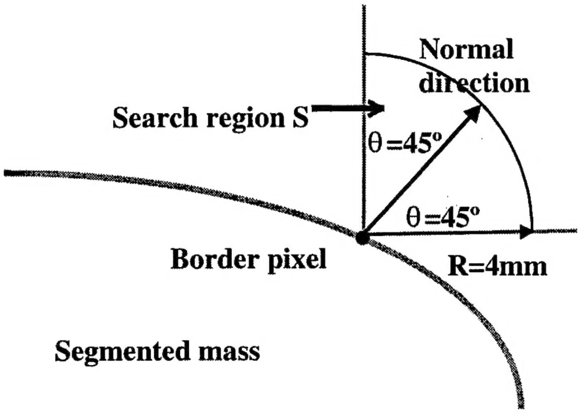


Figure 2. The definition of the search region S

The spiculation likelihood map was thresholded at a constant threshold. The pixels above the threshold were those for which θ was high, i.e., those for which the likelihood of spiculation was high. Three spiculation measures were extracted from the thresholded spiculation likelihood map. These were, i) the number of objects in the thresholded image (NPS), which was related to the number of possible spiculations, ii) the percentage area of the objects (PAS) in the thresholded image relative to the area of the 30-pixel-wide ring, which was related to the percentage area of spiculations, and iii) the product of these two measures (PR).

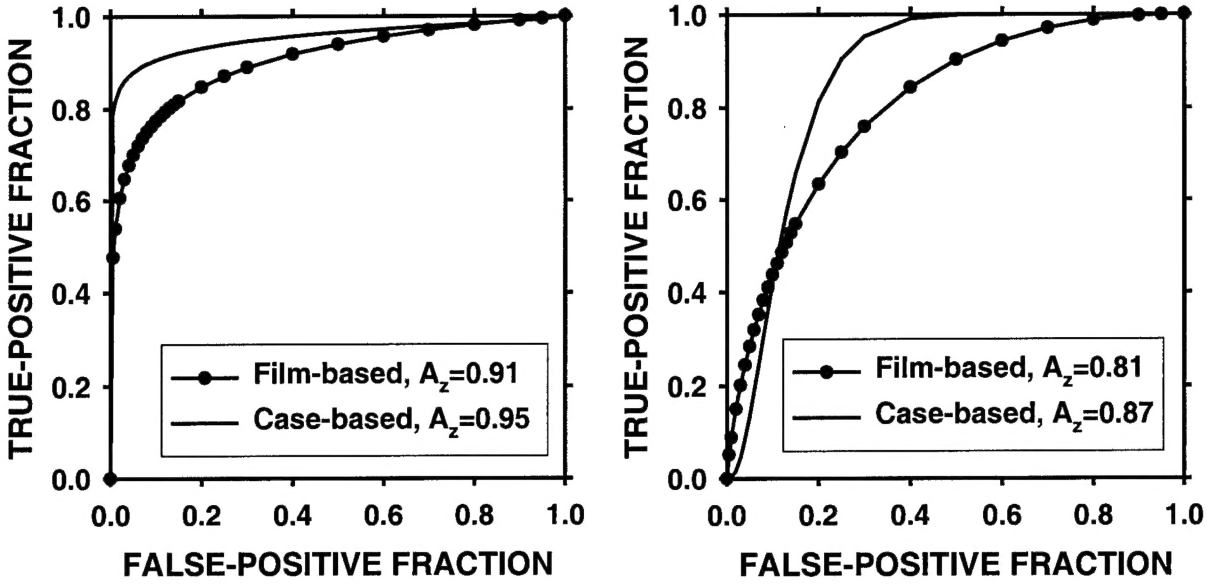
(B) Classification of masses using morphological and spiculation features

These three spiculation measures were used in addition to eleven morphological features extracted from the mass outline for mass characterization. The morphological features were those that were found to be useful for mass characterization in the previous years of our project. The first five morphological features were based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object. These features included NRL mean, standard deviation, entropy, area ratio, and zero crossing count. The remaining six morphological features included the perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast of the object.

The training and test sets used in the evaluation of the classifier were completely independent. Our training data set consisted of 243 mammograms (116 benign and 127 malignant) from 101 patients. Our test data set consisted of 95 mammograms (42 benign and 53 malignant) from 45 patients. A single view was available for nine of these 45 patients. For the remaining 36 test patients, two or more views were available. The true pathology of all the masses was determined by biopsy and histologic analysis.

Stepwise feature selection was used to select effective features for classification from the feature space of fourteen features. Four features, namely, NPS, PR, contrast, and circularity were selected using the set of training regions of interest (ROIs). A backpropagation neural network (BPN) with four input nodes, two hidden-layer nodes, and a single output node was trained using the training set. The accuracy of the designed classifier was evaluated by applying the classifier to test cases that had not been used for training. The test scores were analyzed using receiver operating characteristic (ROC) methodology. The classification accuracy was evaluated as the area A_z under the ROC curve.

We investigated film-based classification of the masses on each mammogram, as well as case-based classification by combining possible multiple views of the same mass. For case-based classification, the BPN scores from different views were averaged. The training A_z values for film-based and case-based classification were 0.91 and 0.95 respectively. The test A_z values for film-based and case-based classification were 0.81 and 0.87. The training and test ROC curves are shown in Figs 3(a) and 3(b), respectively.



(a) (b)

Figure 3 ROC curves for film-based and case-based classification. (a) Training (b) Test.

The difference of this classification method from the independent classification reported last year are the following: 1) The current classification method relies only on morphological and spiculation features, whereas the previous classifier was based on texture and morphological features; and 2) In the current algorithm, we merged information from multiple views of a mass to improve the classification accuracy. A logical next step is to combine spiculation, morphological, and texture features for classification. However, our initial attempts to perform this by using all features in LDA with stepwise feature selection were fruitless. It seems that the spiculation features are very dominant in classification, and once they are selected, the inclusion of texture features actually decreases the classification accuracy. This is a strong indication that more sophisticated classifier is required. We are currently evaluating a hierarchical classifier that will use spiculation features for an initial classification, followed by LDA that uses texture features. We are also continuing to increase our mass database so that both classifier design and testing can be performed with larger data sets.

(C) Feature extraction from computer-extracted microcalcifications

In the third year of the project, we had reported on the development of feature extraction methods from manually identified microcalcifications. In the fourth year of the project, we investigated feature extraction from computer-detected microcalcifications. The ROI to search for the microcalcifications was still manually identified. After the ROI was chosen, the microcalcifications were automatically detected in the ROI containing the cluster. Some of the detections were inevitably false-positives, i.e., non-calcified points that were brighter than their neighboring pixels. In addition, we also had false-negatives, for example, some subtle microcalcifications were not detected by the detection algorithm. Since it will not be possible to manually identify the microcalcifications in practice, the presence of these false-negatives and false-positives represented a more realistic test condition for our characterization algorithms. Since our purpose in this project is lesion characterization, we did not attempt to find the false-positive and false-negative detection rates in this ROIs.

(D) Computer classification for automatically-detected microcalcifications

Starting at the detected locations, the shapes of microcalcifications were extracted using a region growing algorithm. Five morphological features, namely, size, mean density, ratio of second moments, eccentricity of an effective ellipse, and ratio of major and minor axes of the effective ellipse, were extracted from each segmented microcalcification. Since the variations of the shapes and sizes of the individual microcalcifications within a cluster are important for microcalcification classification, the maximum, mean, standard deviation, and coefficient of variation of these individual features were computed for each cluster. The number of microcalcifications in a cluster was also used as a morphological feature. These twenty-one morphological features were the same features that were used in our last yearly report for microcalcification classification.

Four gray level difference statistics features, namely mean, entropy, contrast, and angular second moment were extracted at four different directions from the ROI containing the microcalcification cluster. We thus had 16 texture features.

Texture and morphological features were combined for classification using linear discriminant analysis with stepwise feature selection. The data set for computerized classification consisted of 112 pairs (CC and MLO or CC and LAT) of mammograms. The number of malignant and benign pairs were 40 and 72, respectively. The mammograms were digitized with a Lumisys DIS-1000 laser scanner at a pixel size Of 35mmX35mm and a pixel depth of 12 bits. Leave-one-case-out method was used for both feature selection and classifier parameter estimation. The scores from the two views of a pair were averaged to obtain a score for the pair. Computer classification scores were analyzed by ROC analysis. The accuracy of the classifier was evaluated by the area A_z and the partial area index $A_z(TPF_0)$ above a true-positive fraction of TPF_0=0.90. The computer classifier had an ROC area of 0.83 and a partial area index of 0.42.

(E) Comparison of computer classification and malignancy assessment by radiologists for microcalcifications

We conducted an ROC study in which 7 MQSA-approved radiologists read the same 112 pairs of ROIs. The ROIs were printed on film with a laser printer. The radiologists rated the likelihood of malignancy of each pair on a 10-point rating scale. The case order was randomized for each radiologist. Radiologist ratings were analyzed with ROC methodology. The average ROC curve of 7 radiologists was computed by averaging the slope and intercept parameters of individual ROC curves. The classification accuracy of the radiologists was compared to that of the computer. It was found that the A_z value of the computer was higher than that of all radiologists, and the difference was statistically significant for three of the radiologists (p=0.03). When the partial area index above TPR=0.90 was analyzed, it was found that computer characterization was significantly more accurate than all radiologists ($p<0.05$). The ROC curves and the comparison of the partial area index are shown below.

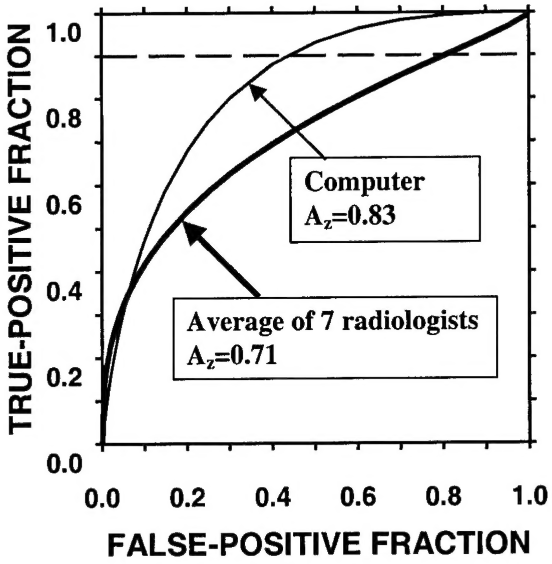


Figure 4 The comparison of the computer classifier and the average of 7 radiologists

	Partial Area Index (TPF>0.9)		Difference	Two-tailed p value
	Computer	Radiologist		
R1	0.42	0.10	0.32	0.0045
R2	0.42	0.05	0.37	0.0008
R3	0.42	0.13	0.29	0.0158
R4	0.42	0.09	0.33	0.0042
R5	0.42	0.06	0.36	0.0018
R6	0.42	0.14	0.28	0.0507
R7	0.42	0.07	0.35	0.0022
Ave. ROC	0.42	0.09	0.33	

Table 1 The comparison of the partial area index (TPF=0.9) between the computer and seven radiologists. The difference between the computer and all seven radiologists was statistically significant.

(7) Appendix

1. Key research accomplishments in current year as a result of this grant

- Features related to the degree of spiculation were extracted from mammographic masses
- A classification algorithm that relies only on morphological and spiculation features was developed
- The accuracy of the mass classification algorithm was tested on a completely independent test set. The combination of this algorithm with texture features still needs to be performed.
- Morphological features were extracted from computer-detected microcalcifications. This is a step toward more realistic implementation of the classification algorithm compared the previous year, in which we had used hand-detected microcalcifications.
- The classification algorithm that was developed in year three was applied to features extracted from computer-detected microcalcifications.
- Using an observer performance study, it was shown that the developed classifier was significantly more accurate than experienced radiologists at the high-sensitivity portion of the ROC curve.

2. Publications in current year as a result of this grant

[1] H.-P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, D.D. Adler, C. Paramagul, J.S. Newman, S.S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiology*, 1999, 212:817-827.

[2] L. Hadjiiski, B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, "Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach", *IEEE Transactions on Medical Imaging*, Vol. 18, No. 12, Dec. 1999, pp. 1178-1187.

[3] H.-P. Chan, B. Sahiner, R.F. Wagner, N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," *Medical Physics,* 1999, 26:2654:2668.

[4] B. Sahiner, H.-P. Chan, N. Petrick, R.F. Wagner, and L.M. Hadjiiski, "Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size," *Proc. SPIE Medical Imaging, 1999,* 3661:499-510.

[5] L.M. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, M.A. Helvie, "Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms," *Proc. SPIE Medical Imaging, 1999,* 3661:464-473.

[6] B. Sahiner, H-P. Chan, N. Petrick, L.M. Hadjiiski, M.A. Helvie, S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," to appear in the proceedings of the International Workshop on Digital Mammography, Toronto, June 2000.

3. Copies of publications are enclosed with this report.

Stepwise linear discriminant analysis in computer-aided diagnosis: the effect of finite sample size

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Robert F. Wagner*, Lubomir Hadjiiski

Department of Radiology, University of Michigan, Ann Arbor, MI 48109-0904
*Center for Devices and Radiological Health, FDA, Rockville, MD 20857

ABSTRACT

In computer-aided diagnosis (CAD), a frequently-used approach is to first extract several potentially useful features from a data set. Effective features are then selected from this feature space, and a classifier is designed using the selected features. In this study, we investigated the effect of finite sample size on classifier accuracy when classifier design involves feature selection. The feature selection and classifier coefficient estimation stages of classifier design were implemented using stepwise feature selection and Fisher's linear discriminant analysis, respectively. The two classes used in our simulation study were assumed to have multidimensional Gaussian distributions, with a large number of features available for feature selection. We investigated the effect of different covariance matrices and means for the two classes on feature selection performance, and compared two strategies for sample space partitioning for classifier design and testing. Our results indicated that the resubstitution estimate was always optimistically biased, except in cases where too few features were selected by the stepwise procedure. When feature selection was performed using only the design samples, the hold-out estimate was always pessimistically biased. When feature selection was performed using the entire finite sample space, and the data was subsequently partitioned into design and test groups, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, number of available samples, and their statistical distribution. All hold-out estimates exhibited a pessimistic bias when the parameters of the simulation were obtained from texture features extracted from mammograms in a previous study.

Keywords: feature selection, linear discriminant analysis, effects of finite sample size, computer-aided diagnosis

1. INTRODUCTION

A common problem in computer-aided diagnosis (CAD) is the lack of a large number of image samples to design a classifier and to test its performance. The effect of finite sample size on the classification accuracy is therefore an important research topic. In order to treat its specific components, previous studies have mostly ignored the feature selection component of this problem, and assumed that the features used in the classifier were fixed.[1-4] However, in many CAD algorithms, feature selection is a necessary first step. This paper addresses the effect of finite sample size on classification accuracy when the classifier design involves feature selection.

In classifier design, the resubstitution and hold-out estimates are commonly used to assess the accuracy of the classifier. To obtain the resubstitution estimate, the classifier is designed using a number of training samples, and the same samples are then applied to the classifier to yield the distribution of the output decision variable for the training group. The resubstitution performance of the classifier is then measured (e.g., by computing the area under the receiver operating characteristic curve, or by evaluating the probability of misclassification) using this distribution. To obtain the hold-out estimate, the classifier is designed in a similar way, except that an independent set of test samples are applied to the classifier to yield the distribution of the output decision variable for the test group. As the number of training samples increases, both of these estimates approach the true classification accuracy, which is the accuracy of a classifier designed with the full knowledge of the sample distributions. When the training sample size is finite, it is known that, on average, the resubstitution estimate of classifier accuracy is optimistic. In other words, it has a higher expected value than the performance obtained with an infinite design sample set. which is the true classification accuracy. Similarly, on average, the hold-out estimate is pessimistic. When classifier design is limited by the availability of design samples, it is important to obtain a conservative (or pessimistic) performance estimate, which provides a lower bound on the classification accuracy.

In CAD literature, different methods have been used to estimate the classifier accuracy when the classifier design involves feature selection. In a few studies, only the resubstitution estimate was provided.[5] In some studies, the researchers partitioned the samples into training and test groups at the beginning of the study, performed both feature selection and

Part of the SPIE Conference on Image Processing • San Diego, California • February 1999
SPIE Vol. 3661 • 0277-786X/99/$10.00

classifier parameter estimation using the training set, and provided the hold-out performance estimate.[6] Several other studies used a mixture of the two methods: The entire sample space was used as the training set at the feature selection step of classifier design, but once the features were chosen, the hold-out or leave-one-out methods were used to measure the accuracy of the classifier.[7-12] To our knowledge, it has not been reported whether this latter method provides an optimistic or pessimistic estimate of the classifier performance.

This paper describes a simulation study that investigates the effect of finite sample size on classifier accuracy when classifier design involves feature selection. We chose to focus our attention on stepwise feature selection in linear discriminant analysis (stepwise linear discriminant analysis) since this is a simple and common feature selection and classification method. The class distributions were assumed to be multivariate Gaussian. We studied the effect of different covariance matrices and means on feature selection performance. We compared the bias of the classifier when feature selection was performed on the entire sample space, and on the design samples alone. The effects of sample size, number of available features, and parameters of stepwise feature selection on classifier bias were examined.

2. METHODS

To evaluate the effect of sample size on feature selection and classifier bias, we studied the problem of stepwise linear discriminant analysis in two stages. The first stage is stepwise feature selection, and the second stage is the estimation of linear discriminant coefficients for the selected feature subset.

2.1. Stepwise Feature Selection

Stepwise feature selection iteratively enters features into or removes features from the group of selected features based on a feature selection criterion.[13] In our study, we used Wilks' lambda, which is defined as the ratio of within-group sum of squares to the total sum of squares of the discriminant scores, as the feature selection criterion. At the feature entry step of the stepwise algorithm, an F value is computed for each feature based on the ratio of the Wilks' lambda before and after the feature is entered into the pool of already selected features. The feature with the largest F value is entered into the selected feature pool if the F value is larger than a threshold F_{in}. At the feature removal step, the features are tested for removal one at a time from the selected feature pool, the F values are computed, and the feature with the smallest F value is removed from the selected feature pool if the F value is smaller than a threshold F_{out}. The algorithm terminates when no more features can satisfy the criteria for either entry or removal. The number of features selected therefore increases, in general, when F_{in} or F_{out} are reduced.

2.2. Estimation of Linear Discriminant Coefficients

As a by-product of the stepwise feature selection procedure used in our study, the coefficients of a linear classifier that classifies its design samples using the selected features are also computed. However, in this study, the design samples used in the stepwise feature selection step of classifier design may be different from those used in the estimation of classifier coefficients. Therefore, we implemented the stepwise feature selection and the classifier coefficient estimation components of our classification scheme separately.

Let Σ_1 and Σ_2 denote the k-by-k covariance matrices of samples belonging to class 1 and class 2, and let $\mu_1 = (\mu_1(1), \mu_1(2), \ldots, \mu_1(k))$ denote their mean vectors. For an input vector X, the linear discriminant classifier output is defined as

$$h(x) = \frac{1}{2}(\mu_2 - \mu_1)^T \Sigma^{-1} X + \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2), \tag{1}$$

where $\Sigma=(\Sigma_1+\Sigma_2)/2$. The linear discriminant classifier is the optimal classifier when the two classes have a multivariate Gaussian distribution with equal covariance matrices.

For the class separation measures considered in this paper (refer to Section 2.3), the constant term $(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)/2$ in Eq. (1) is irrelevant. Therefore, the classifier design can be viewed as the estimation of k parameters of the vector $(\mu_2 - \mu_1)^T \Sigma^{-1}$ using the design samples.

When a finite number of design samples are available, the means and covariances are estimated as the sample means and the sample covariances from the design samples. The substitution of true means and covariances in Eq. (1) by their estimates causes a bias in the accuracy of the classifier. In particular, if the designed classifier is used for the classification of design samples, then the performance is optimistically biased, and if the classifier is used for classifying test samples that are independent from the design samples, then the performance is pessimistically biased.

2.3. Measures of Class Separation

2.3.1. Infinite sample size

When an infinite sample size is available, the class means and covariance matrices can be estimated without bias (i.e., these quantities can be assumed to be known). In this case, we used the Mahalanobis distance $\Delta(\infty)$, or the area $A_z(\infty)$ under the receiver operating characteristic (ROC) curve as measures of classifier accuracy. The infinity sign in parentheses reflects the fact that the distance is computed using the true means and covariance matrices, or, equivalently, using an infinite number of samples.

Assume that the two classes with a multivariate Gaussian distribution with equal covariance matrices have been classified using Eq. (1). Since Eq. (1) is a linear function of the feature vector X, the classifier outputs for class 1 and class 2 will be Gaussian. Let m_1 and m_2 denote means of the classifier output for the normals and the abnormals, respectively, and let s_1^2 and s_2^2 denote the variances for the two classes. With $\Delta(\infty)$ defined as

$$\Delta(\infty)=(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1), \tag{2}$$

it can easily be shown that

$$m_2-m_1=s_1^2=s_2^2=\Delta(\infty). \tag{3}$$

The quantity $\Delta(\infty)$ is referred to as the Mahalanobis distance between the two classes. It is the Euclidean distance between the two classes, normalized to the common covariance matrix.

In particular, if Σ is an k-by-k diagonal matrix with $\Sigma_{i,i}=\sigma^2(i)$, then

$$\Delta(\infty)=\sum_{i=1}^{k}\delta(i), \tag{4}$$

where

$$\delta(i)=[\mu_2(i)-\mu_1(i)]^2/\sigma^2(i) \tag{5}$$

is the squared signal-to-noise ratio of the difference of the means between the two classes for the i^{th} feature.

Using Eq. (3), and the normality of the classifier outputs, it can be shown that[14]

$$A_z(\infty)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{\Delta/2}}e^{-t^2/2}dt \tag{6}$$

2.3.2. Finite sample size

When a finite sample size is available, the means and covariances of the two class distributions were estimated as the sample means and the sample covariances using the training samples, and the classifier outputs for the training and test samples were computed using Eq. (1). The accuracy of the classifier was measured by receiver operating characteristic (ROC) methodology.[15,16] The discriminant scores for samples belonging to class 1 and class 2 were used as decision variables in the LABROC1 program, which provided the ROC curve based on maximum likelihood estimation.

2.4. Simulation conditions

For our simulations, we assumed that the two classes have a multivariate Gaussian distribution with equal covariance matrices, and different means. The number of available features was M=100. We generated a sample size of N_s samples from each class using a random number generator. The sample space was randomly partitioned into N_t training samples and N_s-N_t test samples per class. For a given sample space, we used several different values for N_t in order to study the effect of the design sample size on classification accuracy. In order to reduce the variance of the classification accuracy estimate, a given sample space was independently partitioned 20 times into N_t training samples and N_s-N_t test samples per class, and the classification accuracy using these 20 partitions was averaged. The procedure described above was referred to as an experiment. For each simulation condition described below, 50 statistically independent experiments were performed, and the results were averaged.

Two methods for feature selection were considered. In the first method, the entire sample space was used for feature selection. In other words, the entire sample space was treated as a training set at the feature selection step of classifier design. Before the coefficient estimation step of classifier design, the sample space was partitioned into training and test groups. The training group was used for classifier coefficient estimation, and the resubstitution and hold-out performances were estimated by applying the training and test groups to the designed classifier, respectively. In the second method, sample set partitioning was performed before feature selection. In other words, both feature selection and coefficient estimation were performed only on the training set.

Case 1: Comparison of correlated and diagonal covariance matrices

Case 1.a

In this simulation condition, the 100X100 covariance matrix Σ was chosen to have a block-diagonal structure

$$\Sigma = \begin{bmatrix} A & 0 & 0 & \cdots & 0 \\ 0 & A & 0 & \cdots & 0 \\ 0 & 0 & A & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & A \end{bmatrix}$$

where the 10X10 matrix A was defined as

$$A = \begin{bmatrix} 1 & 0.8 & 0.8 & \cdots & 0.8 \\ 0.8 & 1 & 0.6 & \ldots & 0.6 \\ 0.8 & 0.6 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0.6 \\ 0.8 & 0.6 & \cdots & 0.6 & 1 \end{bmatrix}$$

and $\Delta\mu(i)$=0.1732 for all i. Using (2), the Mahalanobis distance is computed as $\Delta(\infty)$=3.0, and $A_z(\infty)$=0.89.

Case 1.b

The features in Case 1.a can be transformed into a set of uncorrelated features using a linear transformation, which is called the orthogonalization transformation. The linear orthogonalization transformation is defined by the eigenvector matrix of Σ, so that the covariance matrix after orthogonalization is diagonal. After the transformation, the new covariance matrix turns out to be the identity matrix, and the new mean vector is

$$\Delta\mu(i) = \begin{cases} 0.5477 & \text{if } i \text{ is a multiple of 10} \\ 0 & \text{otherwise} \end{cases}$$

Since a linear transformation will not affect the separability of the two classes, the Mahalanobis distance is the same as in Case 1.a, i.e., $\Delta(\infty)$=3.0.

Case 2: Simulation of a possible condition in CAD

In order to simulate covariance matrices and mean vectors that one may encounter in CAD, we used texture features extracted from patient mammograms in a previous study, which aimed at classifying regions of interest (ROIs) on mammograms as malignant or benign.[7] Ten different spatial gray level dependence (SGLD) texture measures were extracted from each ROI at five different distances and two directions. The number of available features was therefore M=100. The transformations that were applied to the ROI before feature extraction, and the formal definition of SGLD features can be found in the literature.[7,17] The means and covariances for each class were estimated from a database of 249 mammograms.

Case 2.a

In this simulation condition, the two classes were assumed to have a multivariate Gaussian distribution with $\Sigma=(\Sigma_1+\Sigma_2)/2$, where Σ_1 and Σ_2 were estimated from the feature samples for the malignant and benign classes. Since the features have different scales, their variances can vary by as much as a factor of 10^6. Therefore, it is difficult to provide an idea about how the covariance matrix is distributed without listing all the entries of the 100X100 matrix Σ. The correlation matrix, which is normalized so that all diagonal entries are unity, is better suited for this purpose. The absolute value of the correlation matrix is shown as an image in Fig. 1. In this image, small elements of the correlation matrix are displayed as darker pixels, and the diagonal elements, which are unity, are displayed as brighter pixels. From Fig. 2, it is observed that some of the features are highly correlated or anticorrelated. The Mahalanobis distance was computed as $\Delta(\infty)=2.4$, which implied $A_z(\infty)=0.86$.

Case 2.b

To determine the performance of a feature space with equivalent discrimination potential, but independent features, we performed an orthogonalization transformation on the SGLD feature space, as explained previously (Case 1.b).

3. RESULTS

Case 1:

Feature selection from the entire sample space

Figs. 2.a and 2.b plot the area A_z under the ROC curve for the resubstitution and hold-out performance estimates versus the inverse of the number of training samples per class, $1/N_t$, for Case 1.a, and Case 1.b, respectively (number of samples per class N_s=100). The F_{in} value was varied between 0.5 and 1.5, and F_{out} was defined as $F_{out}=max[(F_{in}-1),0]$. Fig. 3 is equivalent to Fig. 2.a, except the number of samples per class was increased from N_s=100 to N_s=500 in this figure.

Case 2:

Feature selection from the entire sample space

The area A_z under the ROC curve for the resubstitution and hold-out performance estimates are plotted versus $1/N_t$ in Figs. 4.a and 4.b for Case 2.a, and Case 2.b, respectively (N_s=100). The F_{in} value was varied between 0.5 and 3.0, and F_{out} was defined as $F_{out}=max[(F_{in}-1),0]$. Fig. 5 is equivalent to Fig. 4.a, except the number of samples per class was increased from N_s=100 to N_s=500 in this figure.

Feature selection from training samples alone

Case 2.a was used as an example. The area A_z under the ROC curves versus $1/N_t$ are plotted for N_s=100 and N_s=500 in Figs. 6 and 7, respectively.

4. DISCUSSION

Fig. 2.b demonstrates the potential disadvantage of performing feature selection using the entire sample space. The best possible test performance with infinite sample size for Case 1 is $A_z(\infty)=0.89$. However, in Fig. 2.b, we observe that some of the "hold-out" estimates were as high as 0.92. These estimates were higher than $A_z(\infty)$ because the hold-out samples were excluded from classifier design only in the parameter estimation stage of the design, and were used as training samples in feature selection. When feature selection is performed using a small sample size, some features that are useless for the general population may appear to be useful for the classification of the small number of samples at hand. This was previously demonstrated in the literature by comparing the probability of misclassification based on either a finite sample set or the entire population subject to the constraint that a given number of features were used for classification.[18] In our study, given a small data set, the variance in Wilks' lambda estimates causes some feature combinations to appear more powerful than they actually are. If the data set is partitioned into training and test groups after feature selection, these feature combinations may provide optimistic hold-out estimates.

The observation made in the previous paragraph about feature selection using the entire sample space is not a general rule, however. Figs. 2.a and 4.a show that one does not always run the risk of obtaining an optimistic bias in the hold-out estimate when the feature selection is performed using the entire sample space. For Case 1, the best possible test performance with an infinite sample size is $A_z(\infty)=0.89$, but the best hold-out estimate in Fig. 2.a is A_z=0.82. Similarly, for Case 2, the best possible test performance with infinite sample size is $A_z(\infty)=0.86$, but the best hold-out estimate in Fig. 4.a is A_z=0.84. The features in both Case 1.a and Case 2.a were correlated. Case 1.b and Case 2.b were obtained from Case 1.a and Case 2.a by applying a linear orthogonalization transformation to the features so that they become uncorrelated. Figs. 2.b and 4.b show that after this transformation is applied, the hold-out estimates can be optimistically biased for small sample size

(N_s=100). This shows that performing a linear combination of features before stepwise feature selection can have a dramatic influence on its performance. This result is somewhat surprising, because the stepwise procedure is known to select a set of features whose linear combination can effectively separate the classes. However, the orthogonalization transformation in this study is assumed to be known *a priori* (i.e., it is not deduced from the available finite sample size), and is applied to the entire feature space of M features, whereas the stepwise procedure only produces combinations of a subset of these features.

Figs. 6 and 7 demonstrate that when feature selection is performed using the training set alone, the hold-out performance estimate is pessimistically biased. This bias decreases as the number of training samples, N_t, is increased.

When F_{in} and F_{out} values were low, the resubstitution performance estimates were optimistically biased for all the cases studied. Low F_{in} and F_{out} values imply that many features are selected using the stepwise procedure. From previous studies, it is known that a larger number of features in classification leads to larger resubstitution bias.[3] On the other hand, when F_{in} and F_{out} values were very high, the number of selected features could be so low that the resubstitution estimate would be pessimistically biased, as can be observed from Fig. 3 (F_{in}=1.5) and Fig. 4.a (F_{in}=3.0). In all of our simulations, for a given number of training samples N_t, the resubstitution estimate increased monotonically as the number of selected features were increased by decreasing F_{in} and F_{out}.

In contrast to the resubstitution estimate, the hold-out estimate for a given number of training samples did not change monotonically as F_{in} and F_{out} were decreased. This can be observed from Fig. 2.a, where the hold-out estimate for F_{in}=1.5 is larger than all other hold-out estimates with different F_{in} values for N_t=25 ($1/N_t$=0.04). However, for N_t=90 ($1/N_t$=0.011), the hold-out estimate for the same F_{in} value is no longer the largest. In Fig. 2.a, the feature selection was performed using the entire sample space. A similar phenomenon can be observed in Fig. 7, where the feature selection is performed using the training samples alone. This means that for a given number of design samples, there is an optimum value for F_{in} and F_{out} (or the number of selected features) that provides the highest hold-out estimate. This is the well-known peaking phenomenon described in the literature,[19] which can be explained as follows. For a given number of training samples, increasing the number of features in the classification has two opposing effects on the hold-out performance. On the one hand, the new features may provide some new information about the two classes, which tends to increase the hold-out performance. On the other hand, the same features increase the complexity of the classifier, which tends to decrease the hold-out performance. Depending on the balance between how much new information the new features provide and how much the complexity increases, the hold-out performance may increase or decrease when the number of features is increased.

In this study, the number of available features was fixed at M=100. The number of samples per class was N_s=100 in most of the simulations. However, in three of our simulation conditions, we used N_s=500, which meant that the total number of samples was ten times that of available features. The results of these simulations are shown in Fig. 3 for Case 1, and Figs. 5 and 7 for Case 2. Our first observation concerning these figures is that no hold-out estimates in any of these figures are higher than their respective $A_z(\infty)$ values. This suggests that optimistic hold-out estimates may be avoided by increasing the number of available samples, or, possibly, by decreasing the number of features used for feature selection. A second observation is that, compared to other figures in this study, the relationship between the A_z values and $1/N_t$ is closer to a linear relation. This suggests that it may be possible to obtain $A_z(\infty)$ by fitting a line to the A_z vs. $1/N_t$ curves using linear regression, and finding the y-axis intercept. This is similar to the modified Fukunaga and Hayes technique that we discussed previously in the studies of finite sample size effect on classifier bias.

This study examined only the bias of the mean performance estimates, which were obtained by averaging the estimates from fifty experiments as described in Section 2.4. Another important issue in classifier design is the variance of the individual estimates. The variance provides an estimate of the generalizability of the classifier performance to other design and test samples. We previously studied the variance of performance estimates when the classifier design included the estimation of classifier coefficients, but excluded feature selection.[4,20] The extension of our previous studies to include feature selection is an important further research topic.

5. CONCLUSION

In this study, we investigated the finite-sample performance of a linear classifier that included stepwise feature selection as a design step. We compared the resubstitution and hold-out estimates to the true classification accuracy, which is the accuracy of a classifier designed with the full knowledge of the sample distributions. We compared the effect of partitioning the data set into training and test groups before performing feature selection, and after performing feature selection. When data partitioning was performed before feature selection, the hold-out estimate was always pessimistically biased. When partitioning was performed after feature selection, i.e., the entire sample space was used for feature selection, the hold-out estimates could be pessimistically or optimistically biased, depending on the number of features available for selection, number of available samples, and their statistical distribution. All hold-out estimates exhibited a pessimistic bias when the parameters of the simulation were obtained from correlated texture features extracted from mammograms in our previous study. The understanding of the performance of the classifier designed with different schemes will allow us to utilize a limited sample set efficiently and to avoid an overly optimistic assessment of the classifier

6. ACKNOWLEDGMENTS

This work is supported by a USPHS Grant No. CA 48129 and a USAMRMC grant (DAMD 17-96-1-6254). B. Sahiner and L. Hadjiiski are also supported by Career Development Awards from the USAMRMC (DAMD 17-96-1-6012 and DAMD 17-98-1-8211). N. Petrick is also supported by a grant from the Whitaker Foundation. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC1 programs.

REFERENCES

1. H.-P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: Quadratic and neural network classifiers," Proc. SPIE Conf. Medical Imaging 3034, 1102-1113 (1997).

2. R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Finite-sample effects and resampling plans: Application to linear classifiers in computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging 3034, 467-477 (1997).

3. H.-P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," Proc. SPIE Conf. Medical Imaging 3338, 845-858 (1998).

4. R. F. Wagner, H.-P. Chan, J. Mossoba, B. Sahiner, and N. Petrick, "Components of variance in ROC analysis of CAD_X classifier performance," Proc. SPIE Conf. Medical Imaging 3338, 859-875 (1998).

5. C.-M. Wu, Y.-C. Chen, and K.-S. Hsieh, "Texture feature for classification of ultrasonic liver images," IEEE Transactions on Medical Imaging 11, 141-152 (1992).

6. P. A. Freeborough and N. C. Fox, "MR image texture analysis applied to the diagnosis and tracking of Alzheimer's disease," IEEE Trans. Medical Imaging 17, 475-479 (1998).

7. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. 25, 516-526 (1998).

8. B. S. Garra, B. H. Krasner, S. C. Horri, S. Ascher, S. K. Mun, and R. K. Zeman, "Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis," Ultrasonic Imaging 15, 267-285 (1993).

9. K. G. A. Gilhuijs and M. L. Giger, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," Medical Physics 25, 1647-1654 (1998).

10. M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," IEEE Trans. Medical Imaging 14, 537-547 (1995).

11. Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer," Radiology 187, 81-87 (1993).

12. V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Medical Physics 19, 1475-1481 (1992).

13. N. R. Draper, *Applied regression analysis*, (Wiley, New York, 1998).

14. A. J. Simpson and M. J. Fitter, "What is the best index of detectability," Psychological Bulletin 80, (1973).

15. C. E. Metz, "ROC methodology in radiologic imaging," Invest Radiol 21, 720-733 (1986).

16. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Statistics in Medicine 17, 1033-1053 (1998).

17. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," IEEE Trans. Systems Man Cybernetics SMC-3, 610-621 (1973).

18. S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 252-264 (1991).

19. G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Information Theory 14, 55-63 (1968).

20. R. F. Wagner, H.-P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Components of variance in ROC analysis of CAD_X classifier performance: Applications of the bootstrap," Proc. SPIE Conf. Medical Imaging 3661, (in print) (1999).

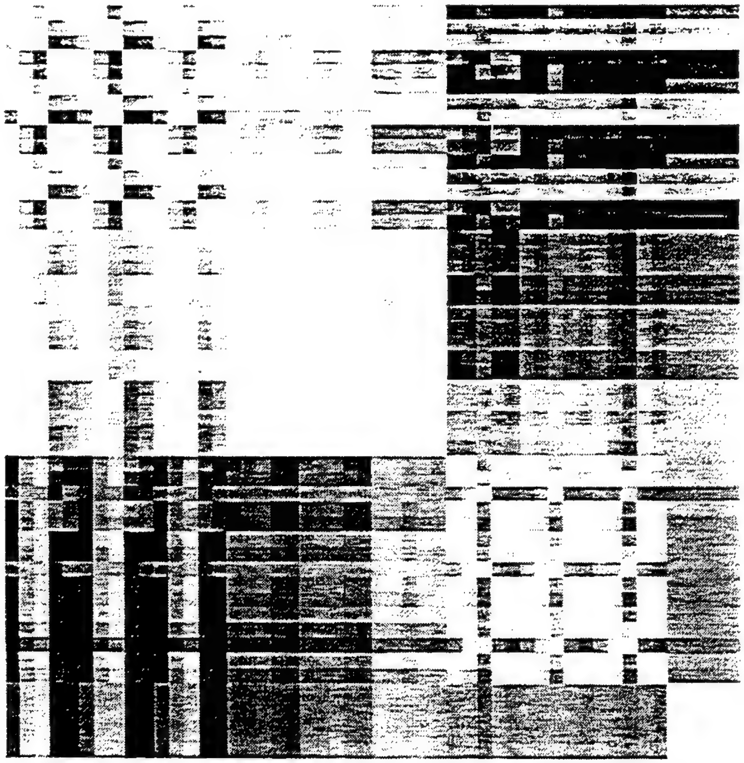


Fig. 1 The absolute value of the correlation matrix for the 100-dimensional texture feature space extracted from 249 mammograms. The covariance matrix corresponding to these features was used in simulation Case 2.a.

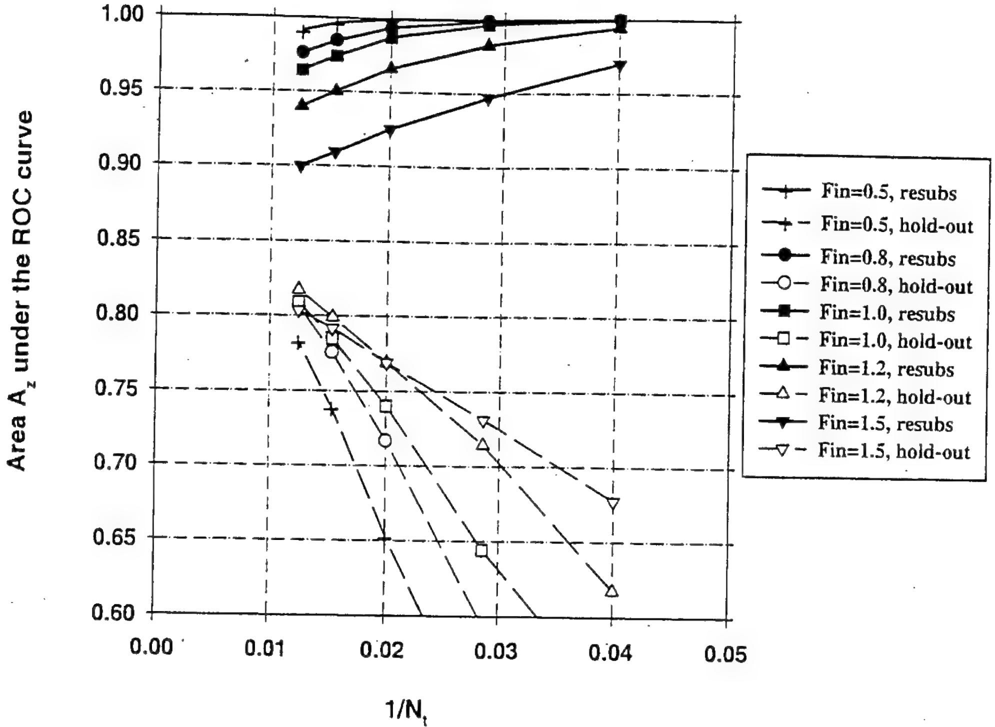


Fig. 2.a The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.a, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

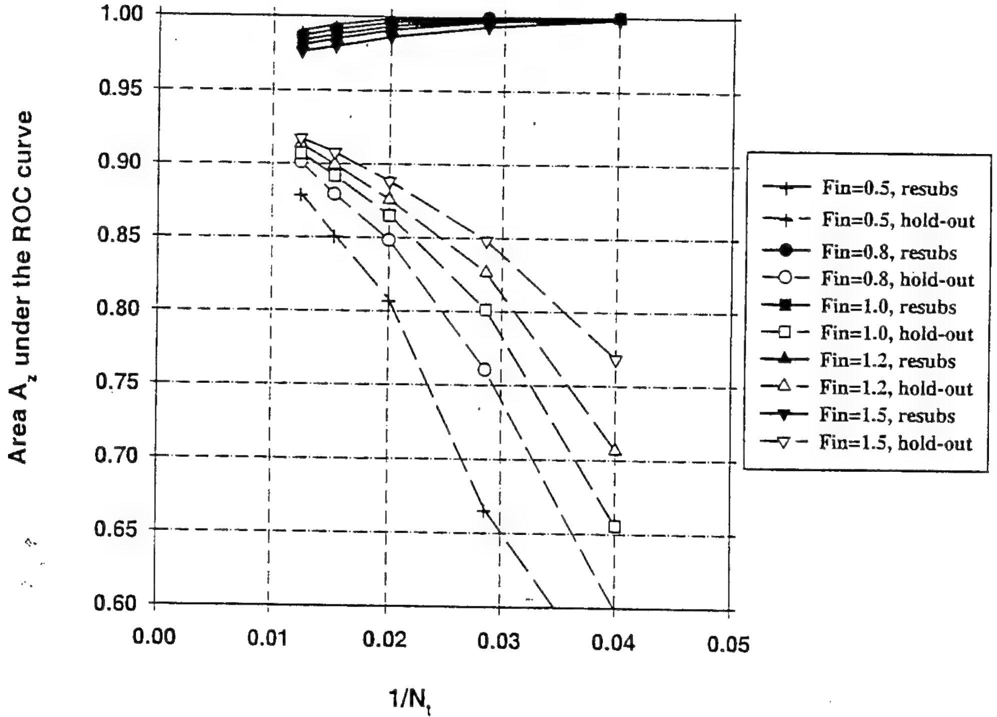


Fig. 2.b The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.b, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

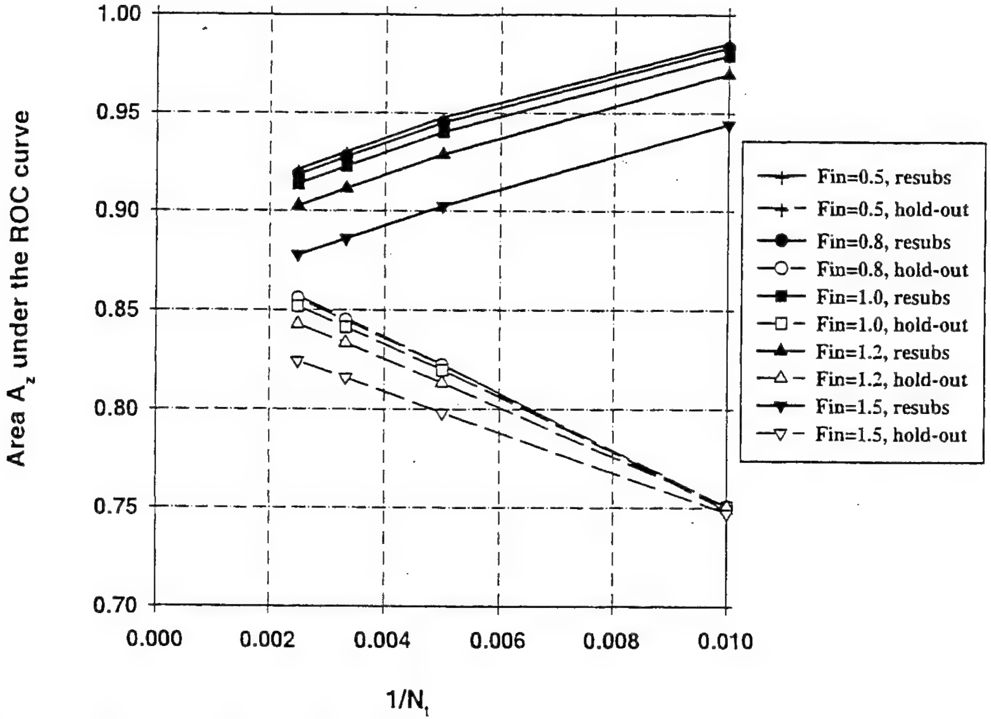


Fig. 3 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 1.a, feature selection from the entire sample space of 500 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.89.

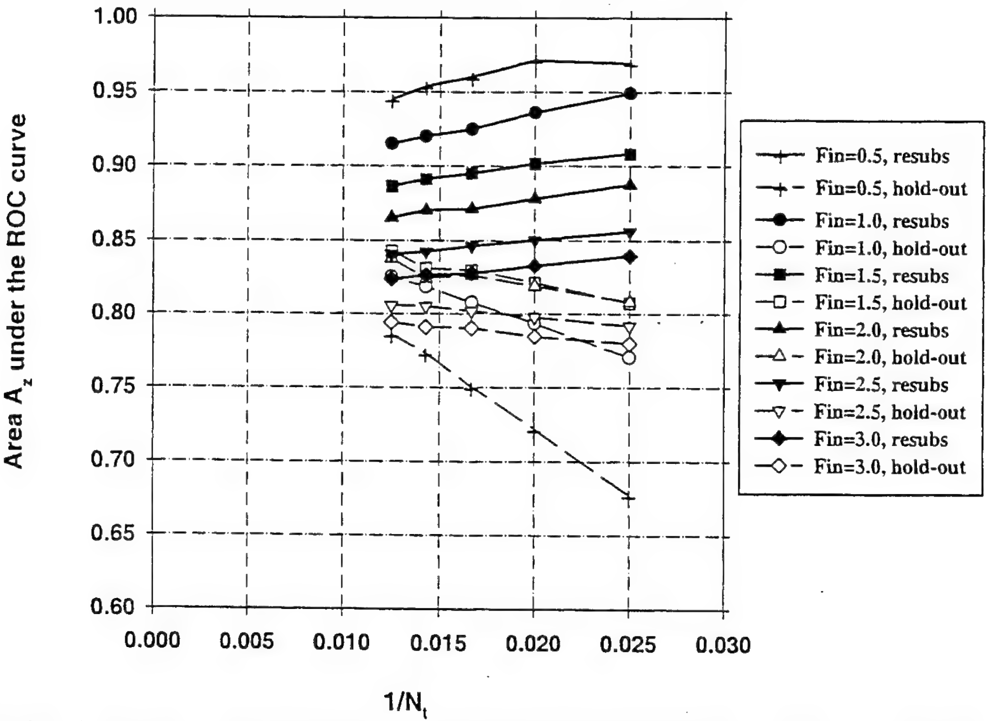


Fig. 4.a The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

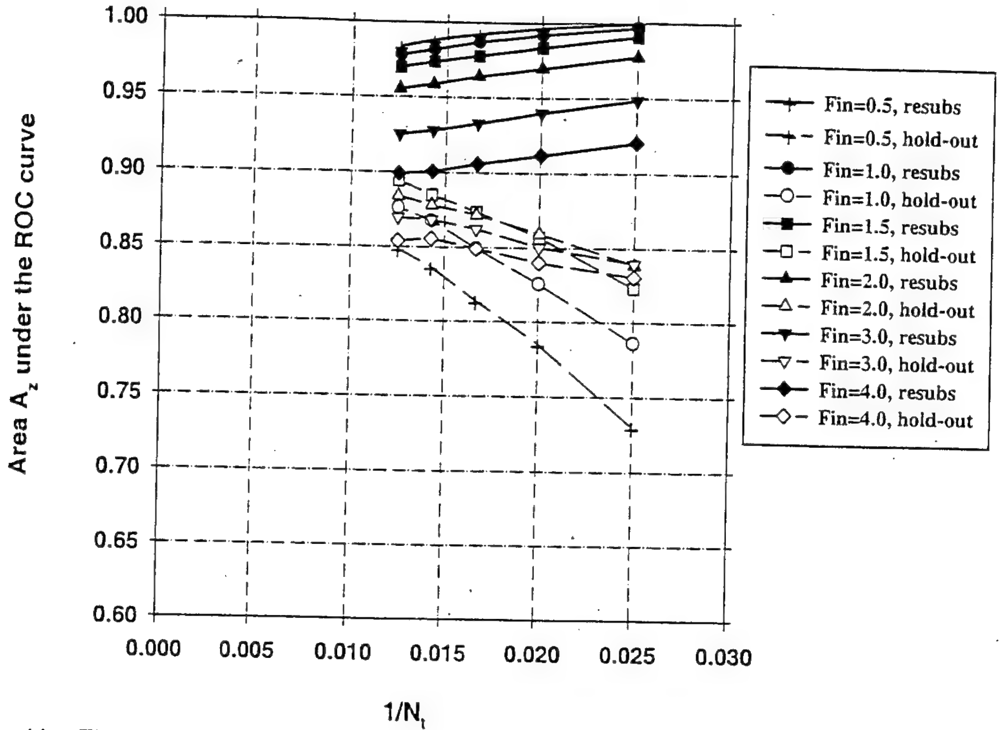


Fig. 4.b The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.b, feature selection from the entire sample space of 100 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

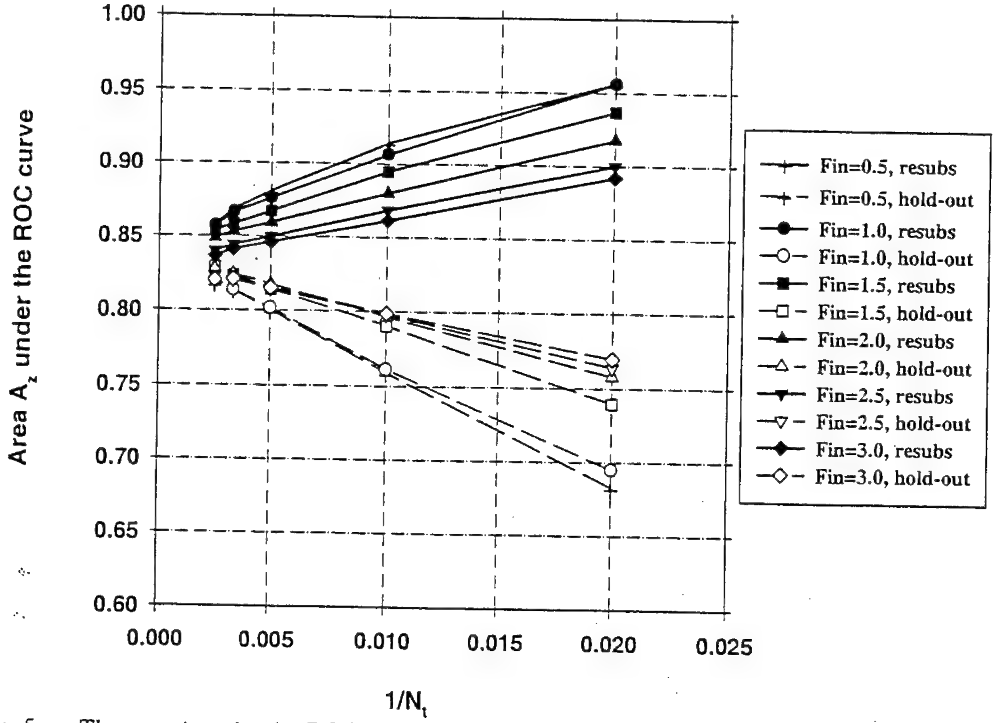


Fig. 5 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from the entire sample space of 500 samples/class. Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

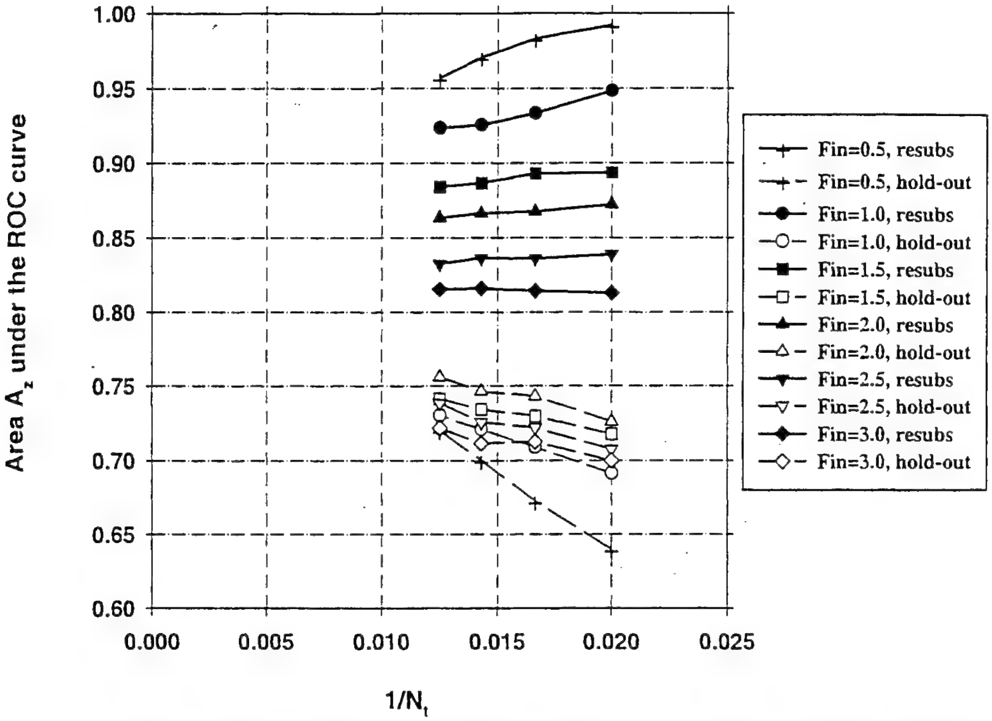


Fig. 6 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from design samples alone (N_s=100). Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

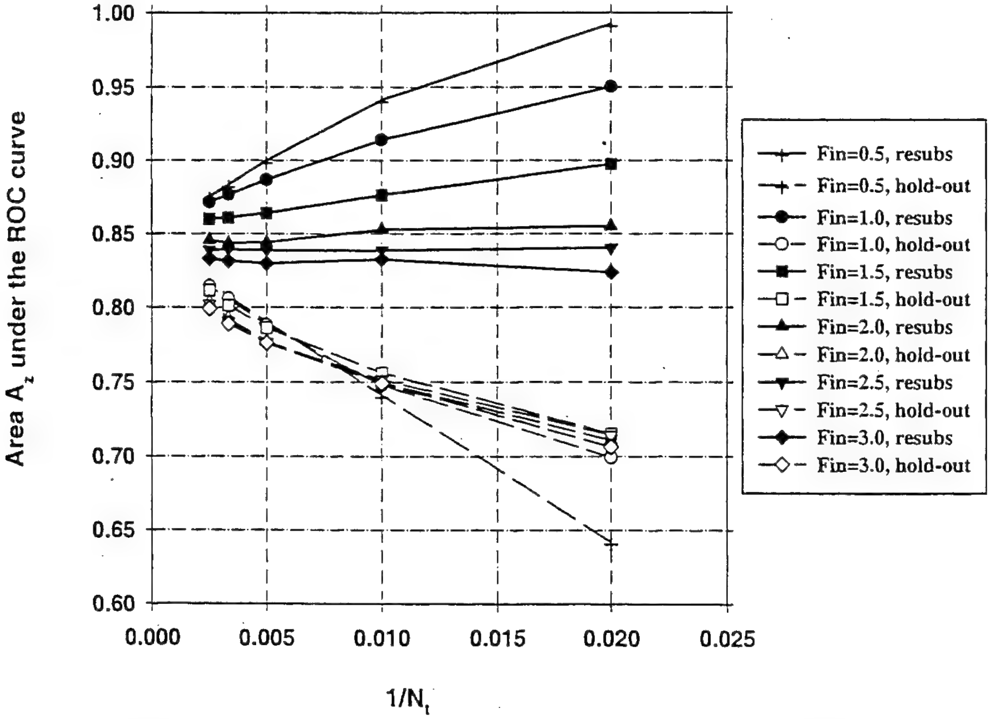


Fig. 7 The area A_z under the ROC curve versus the inverse of the number of design samples N_t per class for Case 2.a, feature selection from design samples alone (N_s=500). Feature selection was performed using an input feature space of M=100 available features. $A_z(\infty)$=0.86.

Hybrid unsupervised-supervised approach for computerized classification of malignant and benign masses on mammograms

Lubomir Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick, Mark Helvie

Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904

ABSTRACT

A hybrid classifier which combines an unsupervised adaptive resonance network (ART2) and a supervised linear discriminant classifier (LDA) was developed for analysis of mammographic masses. Initially the ART2 network separates the masses into different classes based on the similarity of the input feature vectors. The resulting classes are subsequently divided into two groups: (i) classes containing only malignant masses and (ii) classes containing both malignant and benign or only benign masses. All masses belonging to the second group are used to formulate a single LDA model to classify them as malignant and benign. In this approach, the ART2 network identifies the highly suspicious malignant cases and removes them from the training set, thereby facilitating the formulation of the LDA model. In order to examine the utility of this approach, a data set of 348 regions of interest (ROIs) containing biopsy-proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using 73% of ROIs for training and 27% for testing. Classifier design including feature selection and weight optimization was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone. Receiver Operating Characteristics (ROC) analysis was used to evaluate the accuracy of the classifier. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for LDA. The A_z values for the partial areas above a true positive fraction of 0.9 were 0.34 and 0.27 for the hybrid and the LDA classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.

1. INTRODUCTION

Mammography is the most effective method for detection of early breast cancer[1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2-3]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA)[4] and backpropagation neural networks (BPN)[5] which have been shown to perform well in lesion classification problems[6-9]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier[16,17].

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self-organization, decentralization and generalization. It combines the Adaptive Resonance Theory network (ART2)[14,15] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network performs better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events[16,17]. The supervised LDA then classifies the

Part of the SPIE Conference on Image Processing • San Diego, California • February 1999

samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decomposition and data decomposition can improve classification accuracy[10] as well as model accuracy[11]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can deal with the overfitting problem and improve the prediction capabilities of the system.

2. ART2 UNSUPERVISED NEURAL NETWORK

The ART2 is a self organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg[12,13] and a series of further improvements were carried out by Carpenter, Grossberg and co-workers[14,15]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning, i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms such as backpropagation[5] perform slow learning, i.e., they tend to average over occurrences of similar events and require a lot of training iterations.

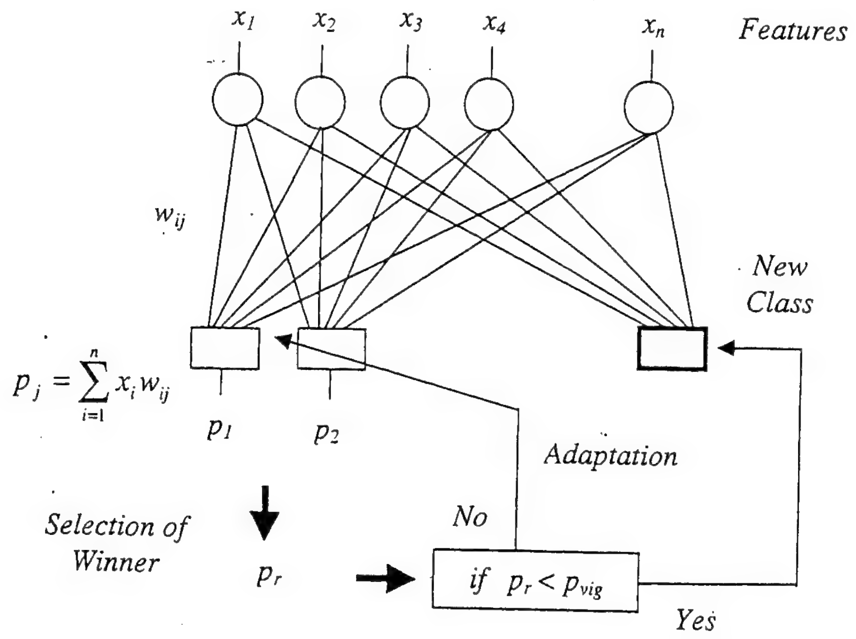


Figure 1. Structure of the ART2 network.

The structure of the ART2 system is shown in Figure 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i (i=1, ... n) and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as:

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \quad j = 1, \dots, k, \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j ($j = 1, \dots, k$) to find the best class match.

Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one[17]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning[14]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{old} + \eta\,(x_i - w_{ir}^{old}) \quad \text{for } i = 1, \dots, n, \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (Eq. 2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the produced activation values for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class ($k+1$) are initialized with the scaled input feature values of this novelty. In this way the activation value p_{k+1} will be maximum ($p_r = p_{k+1}$) and will be higher than p_{vig}, when it is computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between 0 and 1. If p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large the input feature vectors that are more similar will be separated into different classes. The choice of p_{vig} is depends on the particular application.

3. ART2LDA CLASSIFIER

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, these classifiers do not have the ability to correctly classify rare events.

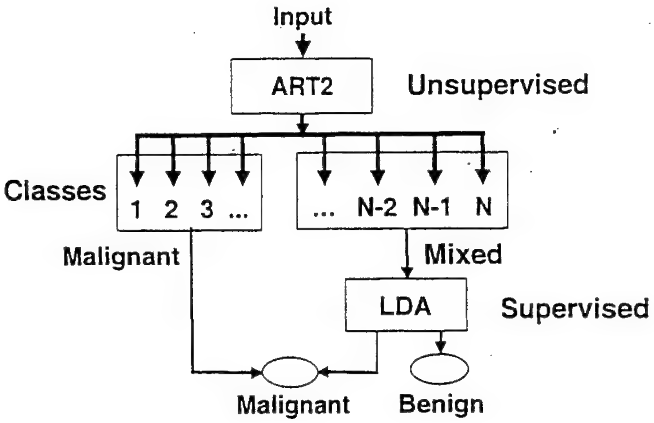


Figure 2. Structure of the ART2LDA classifier.

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 network first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from a multivariate normal distribution for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the normality of the sample distribution by classifying outlying samples into separate classes.

The structure of the hybrid ART2LDA classifier is shown in Fig. 2. The classes identified by ART2 are labeled to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The type of a given class is determined based on ART2 classification of the training data set. The ART2 classifies an input sample into either a malignant or a mixed class. Depending on the class type it is determined whether the LDA classifier will be used. If an input sample is classified into a mixed class, the final classification will be obtained based on the LDA classifier, which has been trained by the mixed classes in the training set. However, if an input sample is classified by ART2 into a malignant class then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor.

4. MATERIALS AND METHODS

4.1. Data set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsy at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. Approximately equal number of malignant and benign masses were included. The data set contained 348 mammograms with a mixture of benign (n=169) and malignant (n=179) masses. The visibility of the masses was rated by a radiologist experienced in breast imaging on a scale of 1 to 10, where the rating of 1 corresponds to the most visible

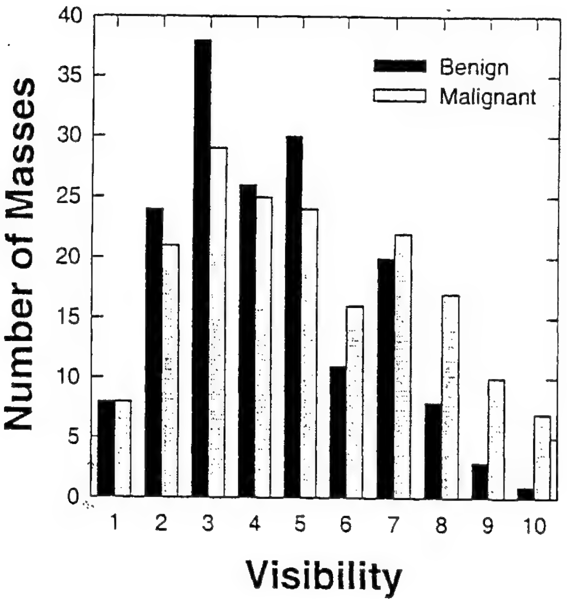


Figure 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced radiologist. (1: very obvious, 10: very subtle).

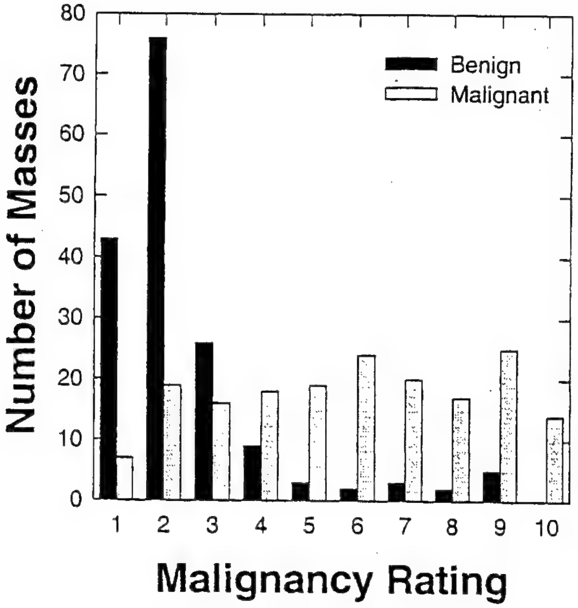


Figure 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced radiologist. (1: very likely benign, 10: very likely malignant).

category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm × 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of –0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0 to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm × 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of –0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were convolved with a 2×2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets. Approximately 73% of the samples have been used for training and 27% for testing. The data set was repartitioned randomly ten times and the training and test results were averaged to reduce their variability.

4.2. Feature extraction

The texture features used in this study were calculated from spatial grey-level dependence (SGLD) matrices[6,7,18] and run-length statistics (RLS) matrices[19]. The SGLD and RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST)[8]. The RBST maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the mass border appears approximately as a horizontal edge, and spiculations appear approximately as vertical lines. A complete description of the RBST can be found in the literature[8].

The (i,j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction 6 at a distance of d pixels apart in an image. Based on our previous studies[6], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12 bit pixel values were discarded. Thirteen texture measures including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances (d=1, 2, 3, 4, 6, 8, 10, 12, 16 and 20) and in four directions (0°, 45°, 90°, and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature[6-8,18]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run[19]. The RLS matrix describes the run length statistics for each gray level in the image. The (i,j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of 5 in the RLS matrix computation could provide good texture characteristics[8].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity, and run percentage were extracted from the vertical and horizontal gradient images in two directions, 6 = 0°, and 6 =90°. Therefore, a total of 20 RLS features were calculated for each ROI.

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

4.3. Feature selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis[20] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda was used as a selection criterion.

4.4. Performance analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program[21], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. The discriminant scores of all case samples classified in the two stages of ART2LDA are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve ($A_z^{(0.9)}$) at a true positive fraction (TPF) higher than 0.9. The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high sensitivity (low false negative) region which is most important for cancer detection in clinical practice.

5. RESULTS

In this study, the test subset was kept truly independent from the training subset; only the training subset was used for feature selection and classifier training, and only the test subset was used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used and the average classification results were estimated.

Table 1. Number of selected features for the 10 data groups.

Data Group No.	1	2	3	4	5	6	7	8	9	10	Mean
Number of selected features	12	15	13	18	14	14	13	18	14	14	14

For a given partition of training and test sets, feature selection was performed based on the training set. The feature selection results for the ten different training groups are shown in Table 1. The average number of selected features was 14. The selected feature sets contained an average of two RLS features and twelve SGLD features. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features.

5.1. ART2LDA classification results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By using different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition no. 3. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone.

In Table 2 the A_z values of the test set for the 10 corresponding partitions are shown. The average test A_z value is 0.81 for the ART2LDA and 0.78 for LDA alone. For nine of the ten partitions, the A_z value was improved by the hybrid classifier.

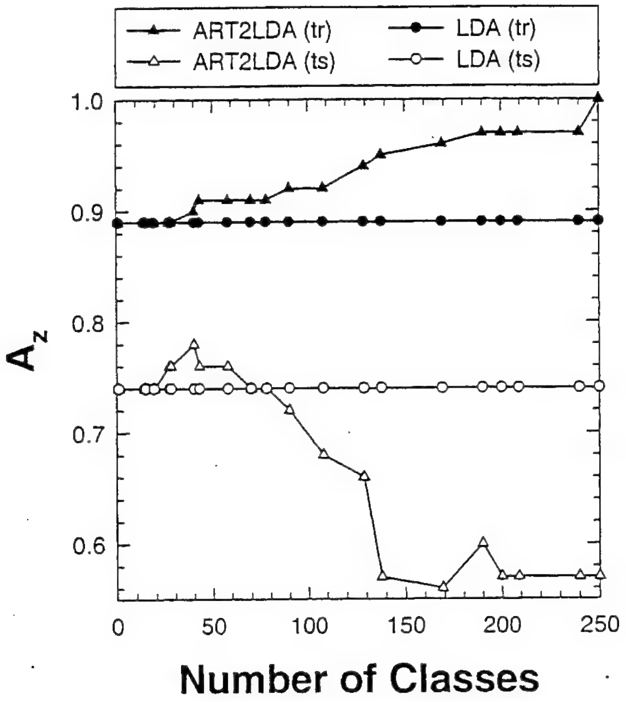


Figure 5. ART2LDA and LDA classification results for training and test sets from data group No.3 as a function of the number of classes generated by ART2.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. In Table 3 the $A_z^{(0.9)}$ values of the test set for the 10 partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high sensitivity operating region (TPF>0.9) of the ROC curve.

Table 2. Classifiers performance for the 10 test sets. The A_z values represent the total area under ROC curve.

Data Group No.	LDA	ART2LDA
1	0.77	0.83
2	0.78	0.80
3	0.74	0.78
4	0.77	0.77
5	0.77	0.78
6	0.80	0.83
7	0.80	0.81
8	0.77	0.80
9	0.77	0.80
10	0.86	0.89
Mean	0.78	0.81

Table 3. Classifiers results for the 10 test sets. The A_z values represent the partial area of the ROC curve above the true positive fraction of 0.9 ($A_z^{(0.9)}$).

Data Group No.	LDA	ART2LDA
1	0.14	0.23
2	0.17	0.21
3	0.19	0.32
4	0.19	0.21
5	0.24	0.26
6	0.27	0.38
7	0.32	0.31
8	0.32	0.34
9	0.40	0.49
10	0.44	0.60
Mean	0.27	0.34

6. DISCUSSION

In this paper a new classifier (ART2LDA) is designed and applied to the classification of malignant and benign masses. The results indicate that the ART2LDA classifier has better generalizability than an LDA classifier alone. The ART2 classifier groups the case samples that are different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depends on how different the outliers are from the rest of the sample population. For the ten different partitions of the training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred tó the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes was generated, an increased number of cases that may be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Fig. 5).

The classification accuracy of ART2LDA increased initially with increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from computational and methodological point of view.

When the partial area of the ROC curve above the true positive (TP) fraction of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers the accuracy of the classification is increased at the high sensitivity end of the curve.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA and the LDA alone, as well as in the differences in the partial $A_z^{(0.9)}$ from the two classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of Student's paired t-test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ (9 out of 10 data set partitions in both cases) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network.

An important difference between the classifier designed in this study and many others in the CAD field is the method of feature selection. In several previously published studies[8,22,23] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was considered to be a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased[24]. In this study, initially the entire data set was partitioned into training and test sets and then feature selection was performed only on the training set. This method results in a pessimistic estimate of the classifier performance[24] when the training set is small. We therefore expect that the performance will be improved when the classifier designed in this study is trained using a large data set. Since our main purpose in this study was to compare the LDA and ART2LDA classifiers, we did not attempt to quantify how pessimistic our results are in this study.

7. CONCLUSION

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of fourteen features were selected for each group. Ten hybrid ART2LDA classifiers and ten LDA models alone were trained by using the ten training sets. The average A_z value under the ROC curve for the test sets was better for ART2LDA (A_z=0.81) compared to the LDA alone (A_z=0.78). A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34 as compared to 0.27 for LDA. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

ACKNOWLEDGMENTS

This work was supported by a USPHS Grant No. CA 48129, and by U.S. Army Medical Research and Materiel Command (USAMRMC) Grant DAMD 17-96-1-6254. Lubomir Hadjiiski is also supported by a Career Development Award from the USAMRMC (DAMD 17-98-1-8211). Berkman Sahiner is also supported by a Career Development Award from the USAMRMC (DAMD 17-96-1-6012). Nicholas Petrick is also supported by a grant from The Whitaker Foundation. The content of this publication does not necessary reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. We would like to thank Prof. Stephen Grosberg and Dr. Gail Carpenter for providing us with valuable information as well as for the useful discussions. Additionally we would like to thank Charles E. Metz, Ph.D. for providing the LABROC1 and CLABROC programs.

REFERENCES

1. H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw-Hill, New York, 1987), pp. 152-172.
2. D. B. Kopans, "The positive predictive value of mammography," *Am. J. Roentgenol.* 158, pp. 521-526, 1992.
3. D. D. Adler, and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.* 4, pp. 123-129, 1992.
4. R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York), 1973.
5. D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart (ed.*), Parallel and Distributed Processing*, Vol. 1, MIT Press, 1986, pp. 318.
6. H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer- Aided Classification of Mammographic Masses and Normal Tissue: Linear Discriminat Analysis in Texture Feature Space," *Phys. Med. Biol.* 40, pp. 857-876, 1995.
7. D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of Mass and Normal Breast Tissue on Digital Mammograms: Multiresolution Texture Analysis," *Med. Phys.*, 22, pp. 1501-1513, 1995.
8. B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized Characterization of Masses on Mamograms: The Rubber Band Sraightening Transform and Texture Analysis," *Med. Phys.* 25 (4), pp. 516-526, April 1998.
9. H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler and M. M. Goodsitt, "Computerized Classification of Malignant and Benign Microcalsifications on mammograms: Texture analysis using an Artificial Neural Network," *Phys. Med. Biol.* 42, pp. 549-567, 1997.
10. M. Jordan, and R. A. Jacobs, "Hierarchical Mixture of Experts and EM Algorithm," *Neural Computation*, 6, pp. 181-214, 1994.
11. L. Hadjiiski, and P. Hopke, "Design of Large Scale Models Based on Multiple Neural Network Approach," *Intelligent Engineering Systems Through Artificial Neural Networks*, Vol. 7, ASME Press, 1997, pp. 61-66.
12. S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biological Cybernetics*, vol.23, no.3, pp.121-134, 1976.
13. S. Grossberg, "Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, illusions," *Biological Cybernetics*, vol.23, no.4, pp. 187-202, 1976.
14. G. A. Carpenter, and S. Grossberg, "ART 2: self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol.26, no.23, 1, pp. 4919-4930, Dec. 1987.
15. G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition," *Neural-Networks*, vol.4, no.4, pp. 493-504, 1991.
16. G. A. Carpenter, and N. Markuzon, "ARTMAP-IC and Medical Diagnosis: Instance Counting and Inconsistent Cases," *Neural-Networks*, vol.11, no.2, pp. 323-336, March 1998.

17. Y. Xie, P. K. Hopke, and D. Wienke, "Airborne Particle Classification with a Combination of Chemical Composition and Shape Index Utilizing an Adaptive Resonance Artificial Neural network," *Environmental Science & Technology*, Vol. 28, No. 11, pp. 1921-1928, 1994.

18. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst. Man Cybern*. 3, pp. 610-621, 1973.

19. M. M. Galloway, "Texture Analysis Using Gray Level Run Length," *Comput. Graph. Image Process*. 4, pp. 172-179, 1975.

20. M. J. Norusis, SPSS Professional Statistics 6.1 (SPSS Inc., Chicago, 1993).

21. C. E. Metz, J. H. Shen, and B. A. Herman, "New Methods for Estimating a Binomial ROC Curve From Continuously Distributed Test Results," *presented at the 1990 Annual Meeting of the American Statistical Association, Anahaim, CA*, 1990.

22. M. F. McNitt-Gray, H. K. Huang, J. W. Sayre, "Feature Selection in the Pattern Classification Problem of Digital Chest Radiograph Segmentation," *IEEE Transaction on Medical Imaging*, Vol. 14, No. 3, pp. 537-547, Sep. 1995.

23. Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated Computerized Classification of Malignant and Benign Masses on Digitized Mammograms," *Acad. Radiol.*, 5, pp. 155-168, 1998.

24. B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *SPIE International Symposium on Medical Imaging*, San Diego, California, February 20-26, 1999., *Proc. SPIE* 3661, (in print).

ACTIVE CONTOUR MODELS FOR SEGMENTATION AND CHARACTERIZATION OF MAMMOGRAPHIC MASSES

Berkman Sahiner, Heang-Ping Chan, Nicholas Petrick,
Lubomir M. Hadjiiski, Mark A. Helvie, Sophie Paquerault
Department of Radiology, University of Michigan, Ann Arbor, MI 48109

Abstract

We have investigated the use of an active contour model for accurate delineation of mass boundaries on mammograms. The model used smoothness constraints and image gradient information in order to refine an initial boundary provided by a clustering algorithm. After segmentation of the mass, possible spiculations were segmented by utilizing gradient direction statistics in a region surrounding the mass. Spiculation measures and morphological features were extracted and used for classifying the mass as malignant or benign. The classification accuracy was evaluated using the area A_z under the receiver operating characteristic (ROC) curve. A data set containing 243 mammograms from 101 patients was used for training the classifier, and a data set containing 95 mammograms from 45 patients were used for testing the classifier. The test A_z for the task of classifying a mass on a single view and a mass on all available views as malignant or benign was 0.81 and 0.87, respectively. Our results indicate that the spiculation measures and the morphological features extracted from automatically segmented mass boundaries are effective in characterizing mammographic masses as malignant or benign.

1. Introduction

In recent years, many researchers have investigated the use of computer-extracted image features for classification of breast masses as malignant or benign (Sahiner *et al.* 1998; Huo *et al.* 1998; Leichter *et al.* 2000). Many features used in computerized breast mass characterization require accurate delineation of mass boundaries as a first step. Accurate computerized delineation of mass boundaries is often difficult because of the presence of ill-defined or obscured boundaries. The human visual system often overcomes this problem by incorporating *a-priori* information, such as smoothness of mass boundaries, with the image information. In order to make use of similar information for computerized mass segmentation, we designed an active contour model based on the image characteristics of mammographic masses. The new model was used to improve the boundaries provided by a clustering algorithm that was developed in our earlier studies. After segmentation, morphological features were extracted from the mass shape, and were combined with spiculation measures for the characterization of breast masses as malignant or benign.

2. Mass segmentation

The location of the biopsied mass was identified by an MQSA-approved radiologist. A region of interest (ROI) containing the biopsied mass was extracted from the mammogram for computerized processing.

2.1. Initial mass segmentation

The mass segmentation method employed in this study started with the initial detection of a mass shape within an ROI using a K-means clustering algorithm. This technique has been discussed in detail in the literature (Sahiner *et al.* 1996). Figures 1(a)-(d) show examples of a spiculated and a nonspiculated mass, and the results of the initial segmentation.

2.2. Active contour segmentation

Although clustering-based mass segmentation resulted in reasonable mass shapes for most of the masses, the segmentation exhibited inaccuracies when the mass was not very conspicuous, or when some parts of the mass were obscured by overlapping normal breast structures. In addition, further refinement was necessary before detection and segmentation of spiculations.

We used an active contour model for the first stage mass shape refinement, and spiculation detection and segmentation for the final shape refinement.

An active contour is a deformable continuous curve, whose shape is controlled by internal forces (the model, or *a-priori* knowledge about the object to be segmented) and external forces (the image). The internal forces impose a smoothness constraint on the contour, and the external forces push the contour towards salient image features, such as edges. To solve a segmentation problem, an initial boundary is iteratively deformed so that the energy due to internal and external forces is minimized along the contour.

The internal energy components in our active contour model were the continuity and curvature of the contour, as well as the homogeneity of the segmented object. The external energy components were the negative of the smoothed image gradient magnitude, and a balloon force that exerted pressure at a normal direction to the contour. The contour was represented by the vertices of an *N*-point polygon whose vertices were $\mathbf{v}(i)=(x(i),y(i)),\ i=1,\ldots,N$. The energy to be minimized was defined as

$$E=\sum_{i=1}^{N}\left[w_{curv}E_{curv}(i)+w_{cont}E_{cont}(i)+w_{grad}E_{grad}(i)+w_{bal}E_{bal}(i)\right]+w_{\hom}E_{\hom} \qquad (1)$$

where each energy term has a weight, *w*.

The curvature energy term is represented by an approximation to the second derivative of the contour, $E_{curv}(i)=\left|\mathbf{v}(i-1)-2\mathbf{v}(i)+\mathbf{v}(i+1)\right|$. This term is large when the angle at vertex *i* is small. By discouraging small angles at vertices, this term attempts to smooth the contour. The continuity term, $w_{cont}E_{cont}(i)$, reflects the deviation of the length of the line segment under consideration from the average line segment length $\bar{d}$. This term favors contours with regular spacing between the vertices over those with irregular spacing. The image gradient magnitude is obtained by smoothing the image with a low-pass filter, finding the partial derivatives in the horizontal and vertical directions, and then computing the magnitude of the partial derivative vector. Since the gradient energy, $E_{grad}(i)$, is defined as the negative of the gradient magnitude, minimizing this term attracts the contour to object edges. The balloon energy encourages the contour to expand in the normal direction, which is required to prevent the contour from collapsing onto itself (Cohen 1991). The purpose of the homogeneity term, $w_{hom}E_{hom}(i)$, is to make the object and the background regions as homogeneous as possible within each region, and to maximize the difference between the two regions (Poon and Braun 1997).

To minimize the contour energy, we used a greedy algorithm that was first proposed by Williams and Shah (Williams and Shah 1992). In this algorithm, the contour was iteratively optimized, starting with the initial contour provided by clustering-based segmentation. At each iteration, a neighborhood of each vertex was examined, and the vertex was moved to the location that minimized the contour energy. Figures 1(c)-(f) show the initial and final contours, respectively, of the model for a spiculated and a nonspiculated mass.

2.3. Segmentation of spiculations

Spiculations on mammograms appear as linear structures with a positive image contrast, and they usually lie in a radial direction to the mass. As a result of their linearity, the gradient directions at image pixels on or close to the spiculation are more or less in the same orientation relative to that of the spiculation. In order to investigate whether a pixel (i_c, j_c) on the mass contour lies on the path of a spiculation, one can make use of this property as follows: In a search region S of the image, compute the statistics of the angular difference θ between the image gradient direction at image pixel (i,j), and the direction of the vector joining pixels (i_c,j_c), and (i,j) (figure 2). If the pixel (i_c,j_c) lies on the path of a spiculation, then θ will be close to $\pi/2$ whenever the image pixel (i,j) is on the spiculation. Therefore, the distribution of θ, obtained from all image pixels (i,j) within the search region S will have a peak around $\pi/2$. If there is no spiculation, and if the gray levels in S are randomly distributed, then this distribution will be uniform. Karssemeijer *et al.* have made use of a similar idea for detecting spiculated lesions on mammograms (Karssemeijer and te Brake 1996), but not for the detection of the actual spiculations. In our method, we combined this idea with the fact that spiculations generally lie in a radial direction to the mass. Therefore, the region S could be limited so that other gradients, such as those resulting from the mass contour itself, can be excluded from the distribution of gradients in S. The details of our spiculation detection method are described in the literature (Sahiner *et al.* 2000; Chan *et al.* 2000). The contours of a spiculated and a nonspiculated mass after spiculation detection are shown figures 1(g) and 1(h), respectively.

3. Feature Extraction and Classification

In the spiculation segmentation stage, three spiculation measures were extracted from each ROI. These were the number of possible spiculations (NPS), the percentage area of spiculations (PAS), and the product of these two measures (PR). These spiculation measures were used in addition to eleven morphological features extracted from the final mass outline for mass characterization. The first five morphological features were based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object. These features included NRL mean, standard deviation, entropy, area ratio, and zero crossing count (Petrick *et al.* 1999). The remaining six morphological features included the perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast of the object. The definition of these features can be found in the literature (Petrick *et al.* 1999).

Stepwise feature selection was used to select effective features for classification from the feature space of fourteen features. Four features, namely, NPS, PR, contrast, and circularity were selected using the set of training ROIs. A backpropagation neural network (BPN) with four input nodes, two hidden-layer nodes, and a single output node was trained using the training set. The accuracy of the designed classifier was evaluated by applying the classifier to test cases that had not been used for training. The test scores were analyzed using receiver operating characteristic (ROC) methodology. The classification accuracy was evaluated as the area A_z under the ROC curve.

4. Data Set

The mammograms used in this study were randomly selected from the files of patients in the Radiology Department at the University of Michigan who had undergone biopsy. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set. Our training data set consisted of 243 mammograms (116 benign and 127 malignant) from 101 patients. Our test data set consisted of 95 mammograms (42 benign and 53 malignant) from 45 patients. A single view was available for nine of these 45 patients. For the remaining 36 test patients, two or more views were available. The true pathology of all the masses was determined by biopsy and histologic analysis.

5. Results

We investigated film-based classification of the masses on each mammogram, as well as case-based classification by combining possible multiple views of the same mass. For case-based classification, the BPN scores from different views were averaged. The training A_z values for film-based and case-based classification were 0.91 and 0.95 respectively. The test A_z values for film-based and case-based classification were 0.81 and 0.87. The training and test ROC curves are shown in figures 3(a) and 3(b), respectively.

6. Discussion and Conclusion

In our previous work, the clustering method was successful in segmenting the main portion of the mass from the background. However, a major limitation of clustering-based segmentation is that, even for well-circumscribed masses, the segmented shape contains many irregularities due to structured or random noises (see figure 1(d)). Another limitation is that, when parts of the mass are obscured by overlapping normal breast structure, clustering method yields inaccurate results. In this study, we used an active contour model for refining the clustering-based segmentation results. By choosing a balance between the active contour weights based on the training set, we were able to obtain object shapes that were mostly smooth, but contours with sharp turns were also possible if the object boundary contained large gradients. Compared to clustering, the resulting boundaries were subjectively judged to be closer to actual mass boundaries. However, the active contour model was not suitable for the segmentation of spiculations. Since the spiculations do not have a large gradient magnitude, the contour cannot have sharp turns at spiculation locations unless w_{curv} is very small. However, a small value for w_{curv} is not practical, because it results in mass shapes that are too irregular all around the contour. For this reason, we designed an additional stage for detection and segmentation of spiculations.

Our results indicate that accurate segmentation of mammographic masses, detection of spiculations, and the use of morphological and spiculation features can be effective in classifying breast masses as malignant or benign.

Acknowledgments

This work is supported by USPHS Grant CA 48129, by a Career Development Award (B.S.) from the USAMRMC (DAMD 17-96-1-6012), and a Whitaker Foundation Grant (N.P.). The content of this publication does not necessarily reflect the position of the funding agencies, and no official endorsement of any equipment and product of any companies mentioned in this publication should be inferred.

References

Chan, H.-P., N. Petrick and B. Sahiner (2000). Computer-aided breast cancer diagnosis. *Soft Computing Techniques in Breast Cancer Prognosis and Diagnosis* Ed. L. C. Jain. New York, CRC Press. (in press).

Cohen, L. D. 1991. On active contour models and baloons. *CVGIP: Img. Underst.* 53: 211-218.

Huo, Z. M., M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt and K. Doi. 1998. Automated computerized classification of malignant and benign masses on digitized mammograms. *Acad. Rad.* 5: 155-168.

Karssemeijer, N. and G. te Brake. 1996. Detection of stellate distortions in mammograms. *IEEE Trans. Med. Img.* 15(5): 611-619.

Leichter, I., S. Fields, R. Nirel, P. Bamberger, B. Novak, R. Lederman and S. Buchbinder. 2000. Improved mammographic interpretation of masses using computer-aided diagnosis. *Eur. Radiol.* 10: 377-383.

Petrick, N., H. P. Chan, B. Sahiner and M. A. Helvie. 1999. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Med. Phys.* 26(8): 1642-1654.

Poon, C. S. and M. Braun. 1997. Image segmentation by a deformable contour model incorporating region analysis. *Phys. Med. Biol.* 42: 1833-1841.

Sahiner, B., H. P. Chan, N. Petrick, M. A. Helvie and M. M. Goodsitt. 1998. Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. *Med. Phys.* 25: 516-526.

Sahiner, B., H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler and M. M. Goodsitt. 1996. Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms. *Med. Phys.* 23: 1671-1684.

Sahiner, B., H.-P. Chan, N. Petrick, M. A. Helvie and L. M. Hadjiiski. 2000. Improvement of mammographic mass characterization using spiculation measures and morphological features. *Med. Phys.* (submitted):

Williams, D. J. and M. Shah. 1992. A fast algorithm for active contours and curvature estimation. *CVGIP: Img. Underst.* 55: 14-26.

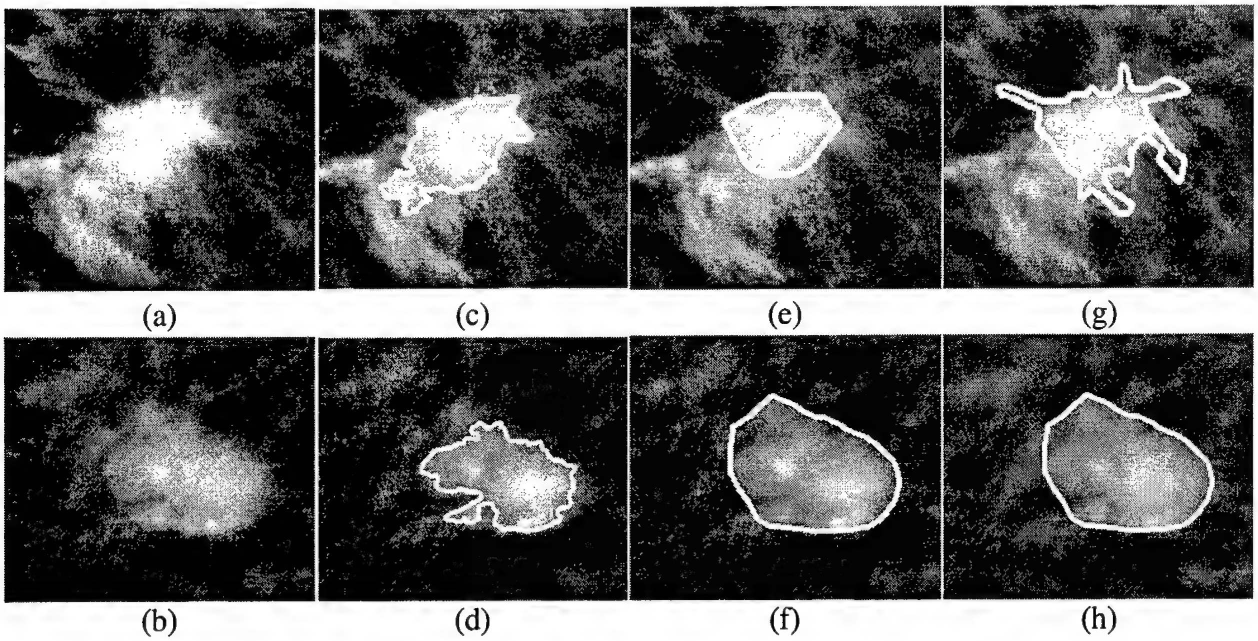


Figure 1. (a), (b) The mass ROI, (c), (d) clustering-based segmentation, (e), (f) active-contour based segmentation, and (g), (h) the result of spiculation detection and segmentation for a spiculated mass (a, c, e, and g) and a nonspiculated mass (b, d, f, and h).

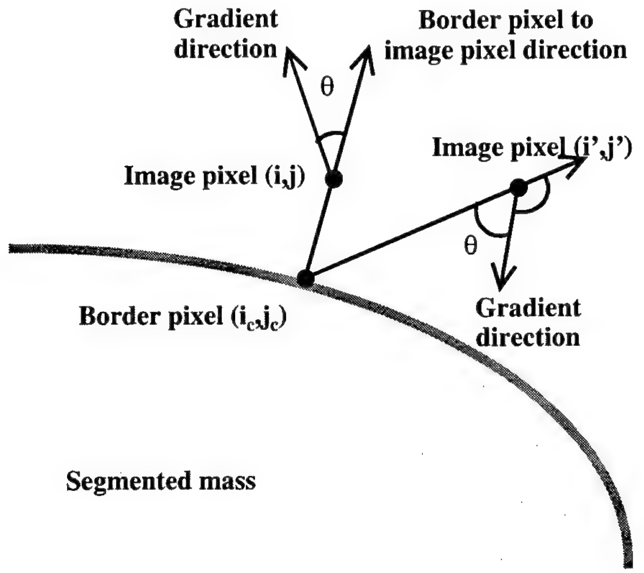
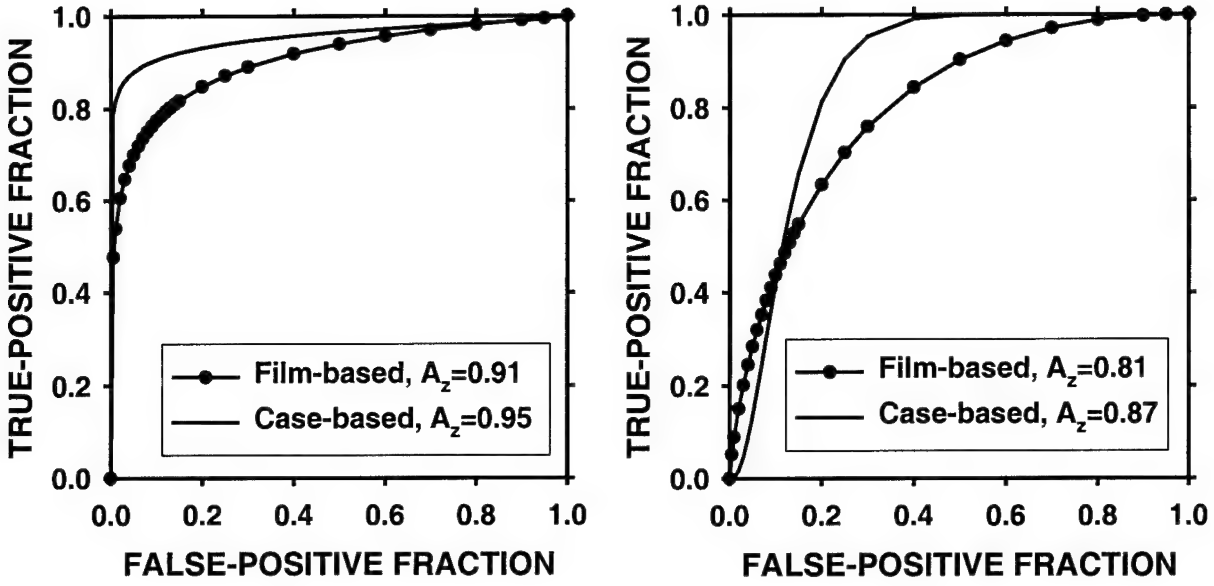


Figure 2. The definition of the angular difference θ.



(a)

(b)

Figure 3. ROC curves for film-based and case-based classification. (a) Training (b) Test.

Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach

Lubomir Hadjiiski,* *Member, IEEE*, Berkman Sahiner, *Member, IEEE*,
Heang-Ping Chan, Nicholas Petrick, *Member, IEEE*, and Mark Helvie

***Abstract*—A new type of classifier combining an unsupervised and a supervised model was designed and applied to classification of malignant and benign masses on mammograms. The unsupervised model was based on an adaptive resonance theory (ART2) network which clustered the masses into a number of separate classes. The classes were divided into two types: one containing only malignant masses and the other containing a mix of malignant and benign masses. The masses from the malignant classes were classified by ART2. The masses from the mixed classes were input to a supervised linear discriminant classifier (LDA). In this way, some malignant masses were separated and classified by ART2 and the less distinguishable benign and malignant masses were classified by LDA. For the evaluation of classifier performance, 348 regions of interest (ROI's) containing biopsy proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using an average of 73% of ROI's for training and 27% for testing. Classifier design, including feature selection and weight optimization, was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone and a backpropagation neural network (BPN). Receiver operating characteristics (ROC) analysis was used to evaluate the accuracy of the classifiers. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for the LDA and 0.80 for the BPN. The partial areas above a true positive fraction of 0.9 were 0.34, 0.27 and 0.31 for the hybrid, the LDA and the BPN classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.**

***Index Terms*— Computer-aided diagnosis, hybrid classifier, mammography, neural networks.**

I. Introduction

MAMMOGRAPHY is the most effective method for detection of early breast cancer [1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2]–[4]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA) [5], [6] and backpropagation neural networks (BPN) [7]–[9] which have been shown to perform well in lesion classification problems [10]–[22]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier [29], [30]. They also have limited generalizability when the training sample set is small.

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self organization, decentralization and generalization. It combines the adaptive resonance theory network (ART2) [26], [27] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network has been found to perform better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events in many classification tasks [28]–[30]. The supervised LDA then classifies the samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decom-

Manuscript received January 27, 1999; revised October 26, 1999. This work was supported by in part by the USPHS under Grant No. CA 48129 and in part by the U.S. Army Medical Research and Materiel Command (USAMRMC) under Grant DAMD 17-96-1-6254. The work of L. Hadjiiski was supported in part by the USAMRMC under Career Development Award DAMD 17-98-1-8211. The work of B. Sahiner was supported in part by the USAMRMC under Career Development Award DAMD 17-96-1-6012. The work of Nicholas Petrick was supported in part by a grant from The Whitaker Foundation. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

*L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, and M. Helvie are with the Department of Radiology, The University of Michigan, Ann Arbor, MI 48109-0904 USA.

Publisher Item Identifier S 0278-0062(99)10410-5.

position and data decomposition can improve classification accuracy [23] as well as model accuracy [24]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can reduce the overfitting problem and improve the prediction capabilities of the system. The performance of the hybrid ART2LDA classifier will be compared with those of an LDA alone or a BPN classifier.

The rest of the paper is organized as follows. In Section II the ART2 unsupervised network is described. A hybrid ART2LDA classifier is introduced in Section III. Section IV describes the data set used in this study. The results are presented in Section V. Section VI contains discussion of these results. Finally, Section VII concludes this investigation.

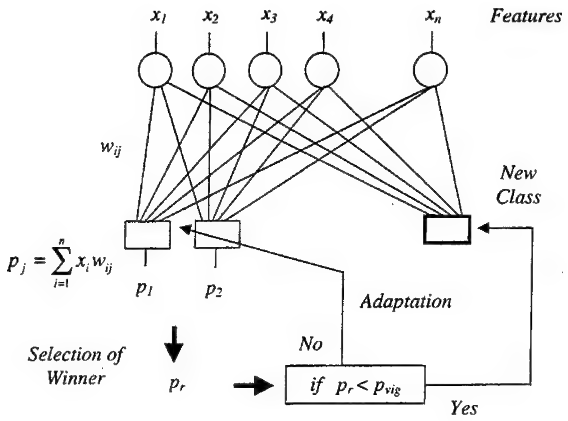


Fig. 1. Structure of the ART2 network.

II. ART2 Unsupervised Neural Network

The ART2 is a self-organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg [25] and a series of further improvements were carried out by Carpenter, Grossberg, and coworkers [26]–[28]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning [28], i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms, such as back propagation [7]–[9], perform slow learning, i.e., they tend to average over occurrences of similar events and require many training iterations.

The structure of the ART2 system is shown in Fig. 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i $(i = 1, \cdots, n)$ and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \qquad j = 1, \cdots, k \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j $(j = 1, \cdots, k)$ to find the best class match. Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one [30]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning [26]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{\text{old}} + \eta(x_i - w_{ir}^{\text{old}}), \qquad \text{for } i = 1, \cdots, n \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the activation values produced for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class $(k + 1)$ are initialized with the scaled input feature values of this novelty. In such a way, the activation value p_{k+1} will be maximum $(p_r = p_{k+1})$ higher than p_{vig} when computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between zero and one. In the case that p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large, the input feature vectors that are more similar will be separated into different classes. The value of p_{vig} is selected differently depending on the particular application.

III. ART2LDA Classifier

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, the learning process in these traditional learning algorithms tends to erase the memory of previous expert knowledge when a new type of expertise is being learned. Therefore, these classifiers do not have as good an ability to correctly classify rare events as ART2 [28], [29].

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from multivariate normal distributions for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the homogeneity of the sample distributions by classifying outlying samples into separate classes.

The ART2LDA hybrid classifier can be described as

$$y_{AL} = g(f_2(x))f_1(x) + 1 - g(f_2(x)) \tag{3}$$

where x is the input vector, $f_1(\cdot)$ is the LDA classifier, $f_2(\cdot)$ is the ART2 classifier, and $g(\cdot)$ is a binary membership function, which labels the classes identified by ART2 to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The membership function is defined as follows:

$$g(c) = \begin{cases} 0, & \text{if } c \text{ is a malignant class} \\ 1, & \text{if } c \text{ is a mixed class.} \end{cases} \tag{4}$$

The type of a given class is determined based on ART2 classification of the training data set.

The structure of the ART2LDA classifier is shown in Fig. 2. The ART2 classifies the input sample x into either a malignant or a mixed class. Depending on the class type the function $g(\cdot)$ determines whether the LDA classifier will be used. If x is classified into a mixed class, the final classification will be obtained based on the LDA classifier. However, if x is classified by ART2 into a malignant class, then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor. This can be seen in (3). The first term in (3), $g(f_2(x))f_1(x)$, is the LDA classifier multiplied by the ART2 control part $g(f_2(x))$. The second term in (3), $(1 - g(f_2(x)))$, gives the classification result of the ART2 stage. If $f_2(x)$ is a malignant class, then $g(f_2(x)) = 0$, the LDA stage is eliminated, and the classifier output y_{AL} is equal to 1. On the other hand, if $f_2(x)$ is a mixed class, then $g(f_2(x)) = 1$, the ART2 term is eliminated, and the final classification is determined by the LDA classifier $(y_{AL} = f_1(x))$.

IV. Methods

A. Data Set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsies at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. The data set contained 348 mammograms with a mixture of benign ($n = 169$) and malignant ($n = 179$) masses. On each mammogram, a region of interest (ROI) containing the mass was identified by a radiologist experienced in breast imaging. The visibility of the masses was rated by the radiologist on a scale of 1 to 10, where the rating of 1 corresponds to the most visible category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

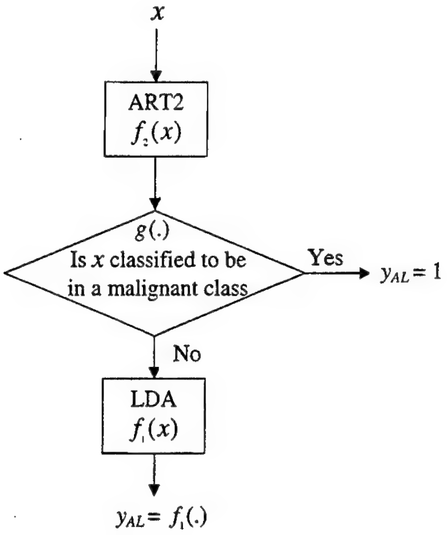


Fig. 2. Structure of the ART2LDA classifier.

The data set can be considered as representative of the patient population that is sent for biopsy under current clinical criteria. Some characteristics of many malignant and benign masses can be visually distinguished by radiologists. However, there is also a nonnegligible fraction of malignant masses that are very similar to benign masses (the low malignancy rating region in Fig. 4). The estimated likelihood of malignancy of malignant and benign masses that are sent for biopsy basically overlaps over the entire range. This is consistent with the fact that in order not to miss malignant masses radiologists must recommend biopsy for even very low suspicion lesions.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm × 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of −0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0

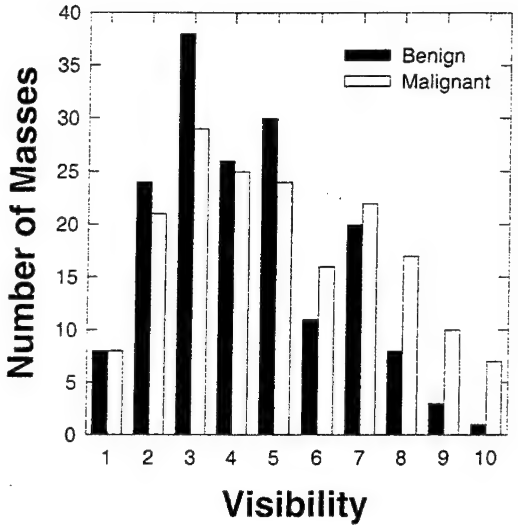


Fig. 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very obvious, 10: very subtle).

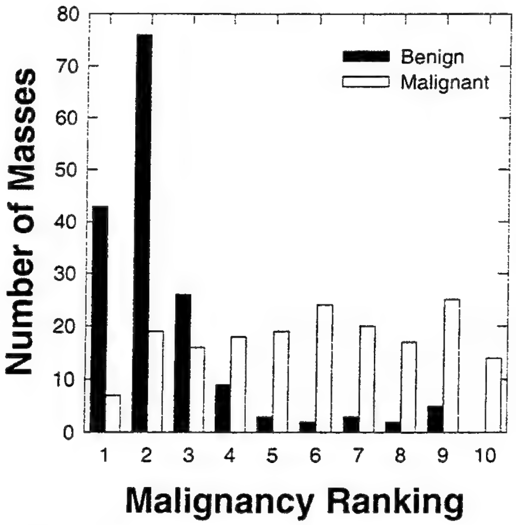


Fig. 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very likely benign, 10: very likely malignant).

to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm × 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of -0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were averaged with a 2 × 2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets on a 3:1 ratio, under the constraints that both the malignant and the benign samples were split with the 3:1 ratio and that the images from the same patient were grouped into the same (training or test) subset. These constraints caused the subsets to deviate from an exact 3:1 ratio. The data set was repartitioned randomly ten times. On average, 73% of the samples were grouped into the training set and 27% into the test set. The training and test results from the ten partitions were averaged to reduce their variability.

B. Feature Extraction

A rectangular ROI was defined to include the radiologist-identified mass with an additional surrounding breast tissue region of at least 40 pixels wide from any point of the mass border. A fully automated method was then used for segmentation of the mass from the breast tissue background within the ROI. The rubber band straightening transform (RBST) was previously developed [12] to map a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the border of mass appears approximately as a horizontal edge and spiculations appear approximately as vertical lines. The transformation of the radially oriented textures surrounding the mass margin to a more uniform orientation facilitates the extraction of texture features.

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices [10]–[12], [31], and run-length statistics (RLS) matrices [32] computed from the RBST images. The (i, j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction at a distance of θ pixels apart in an image. Based on our previous studies [10], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12-bit pixel values were discarded. Thirteen texture measures, including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance, and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances ($d = 1, 2, 3, 4, 6, 8, 10, 12, 16$ and 20) and in four directions (0°, 45°, 90°, and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature [10]–[12], [31]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run [32]. The RLS matrix describes the run length statistics for each gray level in the image. The (i, j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of five in the RLS matrix computation could provide good texture characteristics [12].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity, and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0^\circ$ and $\theta = 90^\circ$. Therefore, a total of 20 RLS features were calculated for each ROI. The formal definition of the RLS feature measures can be found in [32].

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

C. Feature Selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis [33] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda (the ratio of within-group sum of squares to the total sum of squares [34]) was used as a selection criterion. The optimization procedure used a threshold F_{in} for feature entry and a threshold F_{out} for feature removal. On a feature entry step, the features not yet selected are entered into the selected feature pool one at a time, the significance of the change in the Wilks' lambda caused by this feature is estimated based on F statistics. The feature with the highest significance is entered into the feature pool if its significance is higher than F_{in}. On a feature removal step, the features which have already been selected are analyzed one at a time from the selected feature pool and the significance of the change in the Wilks' lambda is estimated. The feature with the least significance is removed from the selected feature pool if the significance is less than F_{out}. Since the appropriate values of F_{in} and F_{out} are not known *a priori*, we examined a range of F_{in} and F_{out} values and chose the appropriate thresholds in such a way that a minimum number of features were selected to achieve a high accuracy of classification by LDA for the training sets. More details about the stepwise linear discriminant analysis and its application to CAD can be found in [10]–[12].

D. Performance Analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology [35]. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program [36], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. For the ART2LDA classifier, the discriminant scores of all case samples classified in the two stages are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA stage .

The performance of ART2LDA was also assessed by estimation of the partial area index ($A_z^{(0.9)}$) and compared with the corresponding performance index of the LDA and BPN classifiers. The partial area index ($A_z^{(0.9)}$) is defined as the area that lies under the ROC curve but above a sensitivity threshold of 0.9 (TPF$_0$ = 0.9) normalized to the total area above TPF$_0$, (1-TPF$_0$). The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high-sensitivity (low false negative) region which is most important for clinical cancer detection task. In addition, the performance of the LDA stage of the ART2LDA classifier was evaluated by the estimation of the area under the ROC curve, denoted as A_z (LDA), for the case samples passed onto the LDA classifier.

TABLE I
NUMBER OF SELECTED FEATURES FOR THE TEN DATA GROUPS WITH THE CORRESPONDING F_{IN} AND F_{OUT} PARAMETERS

Data Group No.	Number of selected features	F_{in}	F_{out}
1	12	1.8	1.6
2	15	2.4	2.2
3	13	2.4	2.2
4	18	2.4	2.2
5	14	2.4	2.2
6	14	2.1	1.8
7	13	2.4	2.2
8	18	1.8	1.6
9	14	2.4	2.2
10	14	2.4	2.2

V. RESULTS

In this section the ART2LDA classification results for malignant and benign masses will be presented and compared with those of the LDA or BPN classifiers. The important point in this study is the fact that the test subset is truly independent of the training subset. Only the training subset is used for feature selection and classifier training, and only the test subset is used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features. The classification result was estimated as the average performance for the ten partitions.

For a given partition of training and test sets, feature selection was performed based on the training set alone. The feature selection results for the ten different training groups are shown in Table I. The average number of selected features was 14. An average of two RLS features and twelve SGLD features were selected for each of the training sets which represented 10% of all RLS features and 2.3% of all SGLD features, respectively. Both types of features (RLS and SGLD) are necessary in order to obtain good classification. The most often selected RLS features for the ten training sets were: horizontal short run emphasis (four times), horizontal long run emphasis (six times), vertical run length nonuniformity (three times), horizontal run length nonuniformity (three times). The most often selected SGLD texture measures for the ten training sets were: inverse difference moment (eight times), information measure of correlations one and two (19 times), difference average (nine times), and correlation (ten times). For a given texture measure, features at different angles or distances may be selected, but these features are usually highly correlated so

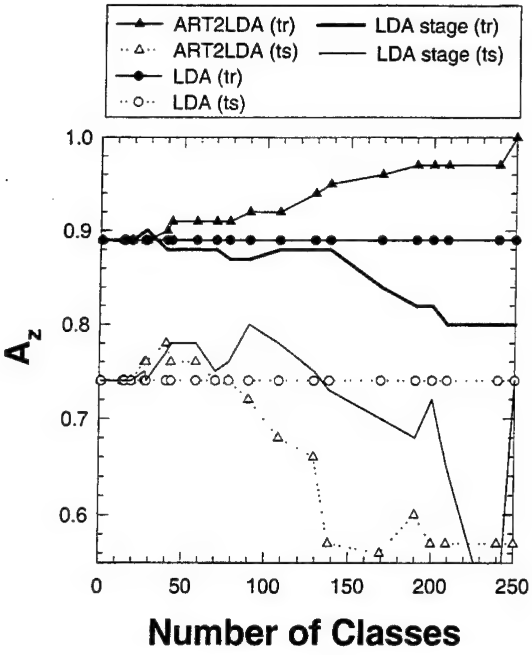


Fig. 5. ART2LDA and LDA classification results for training and test sets from data group three as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

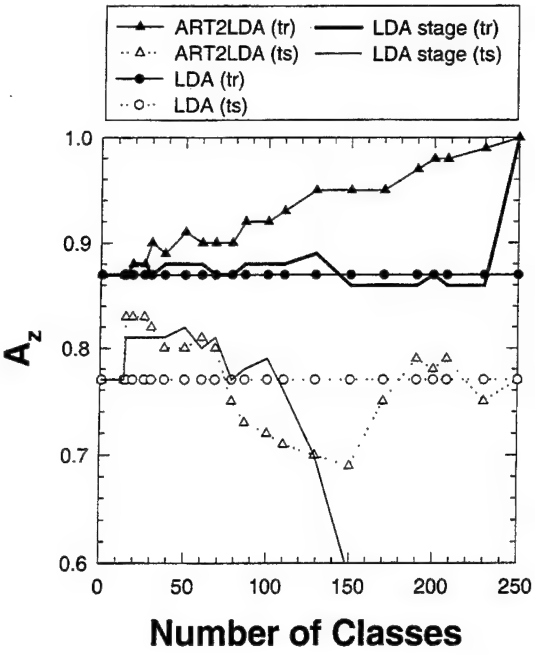


Fig. 6. ART2LDA and LDA classification results for training and test sets from data group one as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

that they can be considered to be similar and counted together as described above.

A. ART2LDA Classification Results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By applying different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2 when p_{vig} was varied. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition three. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone. The two solid lines in Fig. 5 show the A_z values for the LDA stage in the ART2LDA classifier for both the training and test sets. It can be observed that the test A_z for the LDA stage is higher than the A_z for the LDA classifier alone, but not as high as A_z obtained by ART2LDA when the number of classes is small.

In Fig. 6 the classification results of LDA and ART2LDA for the partition one training and test sets are shown. In this case it appeared that in the test set there were two large malignant outliers which degraded the LDA performance. Only 15 classes at the ART2 stage in the ART2LDA was enough to cluster the outliers into a separate malignant class and to improve the performance of the LDA stage and the overall result. The rest of the outliers required more ART2 classes before they were clustered into separate classes and correctly classified as malignant. This is the reason for the similar behavior of the classifiers for partitions three and one in the range of 40 to 70 classes as seen in Figs. 5 and 6. When the number of classes was less than 70, the test A_z for the LDA stage (A_z(LDA)) was higher than the LDA alone, but not as high as the A_z for ART2LDA with less than 30 classes (Fig. 6). The best A_z values for the test data sets of the ten training and test partitions are presented in Table II and Fig. 7. The ART2LDA classifier achieved higher A_z values than the LDA alone in nine of the ten partitions. The average A_z is 0.81 for ART2LDA and 0.78 for LDA alone. The standard deviations of the A_z values for the ten groups range from 0.03 to 0.05 for the ART2LDA classifier and from 0.04 to 0.05 for the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. The results are presented in Table III and Fig. 7. In the lower part of Fig. 7, the $A_z^{(0.9)}$ values of the test set for the corresponding ten partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high-sensitivity operating region (TPF > 0.9) of the ROC curve.

The classifier performance was also evaluated when the ART2LDA classifiers were designed using a fixed number

TABLE II
CLASSIFIERS PERFORMANCE FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE TOTAL AREA UNDER ROC CURVE

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.77	0.83	0.85	0.80
2	0.78	0.80	0.82	0.77
3	0.74	0.78	0.77	0.78
4	0.77	0.77	0.75	0.77
5	0.77	0.78	0.76	0.77
6	0.80	0.83	0.82	0.81
7	0.80	0.81	0.82	0.77
8	0.77	0.80	0.74	0.75
9	0.77	0.80	0.81	0.80
10	0.86	0.89	0.84	0.89
Mean	0.78	0.81	0.80	0.79

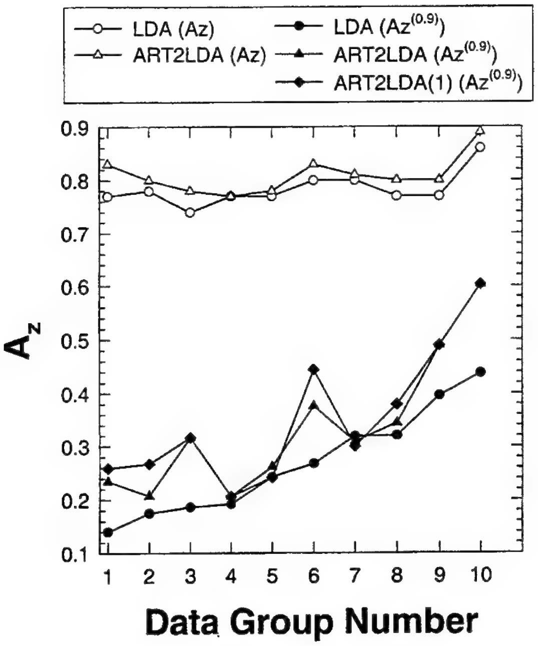


Fig. 7. Average A_z classification results for the 10 test sets. The top graphs represent the ART2LDA and LDA A_z values for the total area under the ROC curve. The bottom graphs represent the ART2LDA, ART2LDA(1) and LDA A_z values for the partial area of the ROC curve above the true positive fraction of 0.9.

TABLE III
CLASSIFIERS RESULTS FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE PARTIAL AREA OF THE ROC CURVE ABOVE THE TRUE POSITIVE FRACTION OF 0.9 ($A_z^{(0.9)}$)

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.14	0.23	0.31	0.26
2	0.17	0.21	0.28	0.27
3	0.19	0.32	0.27	0.32
4	0.19	0.21	0.19	0.21
5	0.24	0.26	0.32	0.24
6	0.27	0.38	0.27	0.44
7	0.32	0.31	0.38	0.30
8	0.32	0.34	0.25	0.38
9	0.40	0.49	0.40	0.49
10	0.44	0.60	0.38	0.60
Mean	0.27	0.34	0.31	0.35

TABLE IV
AVERAGE A_z AND AVERAGE $A_z^{(0.9)}$ CLASSIFICATION RESULTS FOR THE TEN TEST SETS. CLASSIFIERS WERE DESIGNED USING A FIXED NUMBER OF ART2 CLASSES

	LDA	ART2LDA					
No. of classes		15	20	30	40	50	60
A_z	0.78	0.80	0.80	0.80	0.80	0.78	0.77
$A_z^{(0.9)}$	0.27	0.30	0.31	0.33	0.33	0.31	0.31

of ART2 classes. The A_z, and $A_z^{(0.9)}$ results, averaged over the ten test partitions, are presented in Table IV. The average A_z with the ART2LDA classifier, compared to that of LDA alone, was again improved between 15 and 40 classes. The maximum average A_z of 0.80 was achieved between 20 and 40 classes. The average $A_z^{(0.9)}$ results are improved for all ART2LDA classifiers presented in Table IV. The maximum average $A_z^{(0.9)}$ value is 0.33 and it remains constant between 30 and 40 classes.

An alternative way to evaluate the performance of a classifier is its classification accuracy when a decision threshold for malignancy is selected based on the training set. For instance, a decision threshold may be selected such that all positive samples from the training set are classified correctly i.e., at a sensitivity of 100%. The ART2LDA with this decision threshold is referred to as ART2LDA(1). For a given training and test partitioning, ART2LDA classifiers with different number of classes in the ART2 stage were obtained (Figs. 5 and 6). For each of these models the decision threshold for a sensitivity of 100% was selected from the training set and the corresponding ART2LDA(1) classifier was obtained. Then the ART2LDA(1) classifier (with a specific number of classes in the ART2 stage) that correctly classified the maximum number of malignant masses in the test set is selected. By using all samples of the test set, the A_z value is calculated for the corresponding ART2LDA model. The A_z values for the ART2LDA(1) classifiers for the test sets of the ten data partitionings are shown in Tables II and III. For five of the partitions the overall A_z value for ART2LDA(1) is higher than that of LDA alone (Table II). The average A_z value was 0.79. The partial areas above the TP fraction of 0.9, $A_z^{(0.9)}$, for the ten test data sets obtained by the ART2LDA(1) classifier are also shown in Fig. 7. The ART2LDA(1) achieved the highest average $A_z^{(0.9)}$ value of 0.35 compared to ART2LDA and LDA (Table III).

B. BPN Classification Results

A multilayer perceptron back-propagation neural network with a single hidden layer and a single output node was used for comparison with the ART2LDA classifier. The number of selected features determined the number of input nodes to the BPN. The same ten training/test set partitions (as in the case of ART2LDA) were used for the training and validation of the BPN classifiers. BPN's with their number of hidden nodes ranging from two to ten were evaluated to obtain the best architecture. Back-propagation training was used. Each of the BPN's was trained for up to 18 000 training epochs. At every 1000 epochs the neural network weights were saved and the classification result for the corresponding test set was evaluated. This design procedure was repeated for each of the ten training/test groups. For each group, the best test result among all the BPN architectures (different number of hidden nodes) and all the training epochs examined was selected. The average test A_z over the ten groups for the BPN was 0.80, compared to 0.81 for ART2LDA (Table II). The standard deviations of the A_z values for the ten groups range from 0.04 to 0.05 for the BPN. The average partial $A_z^{(0.9)}$ for the BPN was 0.31, compared to 0.34 for ART2LDA (Table III). The A_z and $A_z^{(0.9)}$ of the ART2LDA classifier were higher than those of the BPN in six of the ten training/test groups.

VI. Discussion

In the present study, a new classifier (ART2LDA) was designed and applied to the classification of malignant and benign masses. The results indicated that the ART2LDA classifier had better generalizability than an LDA classifier alone. The ART2 classifier grouped the case samples that were different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depended on how different the outliers were from the rest of the sample population. For the ten different partitions of training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred to the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes were generated, an increased number of cases that might be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Figs. 5 and 6).

The classification accuracy of ART2LDA increased initially with an increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second-stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from a computational and a methodological point of view.

For the optimal number of classes (usually less than 50 for the data sets used) the A_z value for the second-stage LDA in the ART2LDA was better than an LDA classifier alone, but it was not as good as the overall A_z from the ART2LDA. It is evident that the ART2 was a useful classifier for improvement of the second-stage classification.

When the partial area of the ROC curve above the true positive fraction (TPF) of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers, the accuracy of the classification was increased at the high sensitivity end of the curve.

The classifier performance was evaluated when the ART2LDA classifiers were designed using a fixed number of ART2 classes. The results showed improved performance of the ART2LDA in a range between 20 and 40 ART2 classes. Both the average A_z and the average $A_z^{(0.9)}$ reached a maximum within this region, and the maximum average A_z and the average $A_z^{(0.9)}$ values remained unchanged between 30 and 40 classes. These results indicated that the performance of a hybrid ART2LDA classifer was robust and stable and could be potentially useful in real clinical applications.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA, the LDA alone, and the BPN, as well as in the differences in the partial $A_z^{(0.9)}$ from the three classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of student's paired t test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ by the ART2LDA (9 out of 10 data set partitions in both cases for LDA and six out of ten data set partitions in both cases for BPN) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network. In addition, one advantage of the ART2LDA is that the training process is more efficient than that of the BPN, especially when there is a subset of outlying samples. In such a case, the BPN will require a large number of training epochs to minimize the error function.

ART2LDA can be trained to classify the sample cases into more than two classes, such as a class of normal tissue regions in addition to malignant and benign masses. There will be an increase in the complexity of training and a larger training sample size will be desired, but these requirements will be comparable for the different classifiers. In a clinical situation, if the classification task is performed on all computer-detected lesions, the classifier has to distinguish the falsely detected normal tissue from malignant or benign lesions. However, it may be noted that a classifier that can distinguish only malignant and benign masses is applicable to the scenario that the radiologist identifies a suspicious lesion on the mammogram and would like to have a second opinion about its likelihood of malignancy before making a diagnostic decision. Therefore, the development of a classifier that can differentiate malignant and benign masses is the research of interest for many investigators.

Similarly, ART2 can be trained to discover and remove a pure benign mass class. The approach will be similar to the task of classifying and removing the pure malignant classes, as described in this study. However, our approach of removing the malignant classes will reduce the chance of misclassification of malignant masses. In breast cancer detection, the cost of false-negative (missed cancer) is very high. Therefore, our goal in classifier design is to be conservative. By removing the malignant classes in the first stage, any misclassification to these classes will be regarded as malignant. The remaining classes will be classified again with the second-stage classifier so malignant masses will be less likely to be missed.

The problem of classification of malignant and benign masses has been studied by many investigators. Rangayyan *et al.* [15] used Mahalanobis distance classifer (a modification of an LDA classifier) and the leave-one-out method to evaluate the classification of 54 masses. Fogel *et al.* [16] compared LDA and BPN classifiers using the leave-one-out method and 139 masses (malignant and benign classification). Highnam *et al.* [17] used a morphological feature called a halo to classify 40 masses as malignant and benign. Huo *et al.* [22] employed BPN and a rule-based classifier to classify 95 masses using the leave-one-out evaluation method. Sahiner *et al.* [12] used an LDA classifier and the leave-one-out method to classify 168 masses. An important difference between the classifier designed in this study and the previous studies in the CAD field is the method of feature selection. In the above mentioned studies [12], [15]–[17], [22] and several other published studies [18]–[21] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was used as a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased [37]. In our current study, the entire data set was initially partitioned into training and test sets and then feature selection was performed only on the training set. This method will result in a pessimistic estimate of the classifier performance when the training set is small [37]. However, it will provide a more conservative but realistic estimation of the classifier performance in the general patient population. We can expect that the performance would be improved if the classifier in this study were designed using a large data set. Since our main purpose in this study was to compare the ART2LDA classifier with the commonly used LDA and BPN, we did not attempt to quantify how pessimistic our results were in this study.

The most important contribution of this paper is to introduce a new approach that utilizes a two-stage unsupervised–supervised hybrid classifier. We believe that the hybrid approach will improve classification when the sample distribution contains subpopulations that may be difficult for a single classifier to classify. It will be useful for similar classification tasks although different classifiers may be used in each stage of the hybrid structure.

VII. Conclusion

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of features were selected for each group. A hybrid ART2LDA classifier, an LDA, and a BPN were trained by using each of the ten training sets. The A_z value under the ROC curve for the test sets, averaged over the ten partitions, was higher for ART2LDA ($A_z = 0.81$) compared to those of the LDA alone ($A_z = 0.78$) and of the BPN ($A_z = 0.80$). A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34, as compared to 0.27 for LDA and 0.31 for BPN. Additionally, for the ART2LDA classifiers that correctly classified the maximum number of malignant masses in the test sets with decision threshold defined with the training set, the average partial A_z was 0.35. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

Acknowledgment

The authors would like to thank Prof. S. Grosberg and Dr. G. Carpenter for providing them with valuable information as well as for the useful discussions. Additionally the authors would like to thank C. E. Metz, Ph.D., for providing the LABROC1 and CLABROC programs.

References

[1] H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, I. M. Ariel and J. B. Cleary, Eds. New York: McGraw-Hill, 1987, pp. 152–172.

[2] D. B. Kopans, "The positive predictive value of mammography," *Amer. J. Roentgenol.*, vol. 158, pp. 521–526, 1992.

[3] D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.*, vol. 4, pp. 123–129, 1992.

[4] M. Moskowitz, "Impact of a priory medical detection on screening for breast cancer," *Radiology*, vol. 184, pp. 619–622, 1989.

[5] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner, 1975.

[6] R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[7] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1974.

[8] D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart, Ed., *Parallel and Distributed Processing*. Cambridge, MA: MIT Press, 1986, vol. 1, p. 318.

[9] J. Herz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.

[10] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminat analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857–876, 1995.

[11] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Med. Phys.*, vol. 22, pp. 1501–1513, 1995.

[12] B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mamograms: The rubber band sraightening transform and texture analysis," *Med. Phys.*, vol. 25, no. 4, pp. 516–526, Apr. 1998.

[13] B. Sahiner, H. P. Chan, D. Wei, N. Petick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Med. Phys.*, vol. 23, no. 10, pp. 1671–1683, Oct. 1996.

[14] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant

and benign microcalsifications on mammograms: Texture analysis using an artificial neural network," *Phys. Med. Biol.*, vol. 42, pp. 549–567, 1997.

[15] R. M. Rangayyan, N. M. El-Farmawy, J. E. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans. Med. Imag.*, vol. 16, pp. 799–810, Dec. 1997.

[16] D. B. Fogel, E. C. Wasson, E. M. Boughton, V. W. Porto, and P. J. "Angeline, linear and neural model for classifying breast masses," *IEEE Trans. Med. Imag.*, vol. 17, pp. 485–488, June 1998.

[17] R. P. Highnam, J. M. Brady, and B. J.Shepstone, "A quantitative feature to aid diagnosis in mammography," in *Proc. Digital Mammography'96*, pp. 201–206.

[18] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[19] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvements in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," *Med. Phys.*, vol. 19, pp. 1475–1481, 1992.

[20] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *IEEE Trans. Med. Imag.*, vol. 12, pp. 664–669, Dec. 1993.

[21] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," *IEEE Trans. Med. Imag.*, vol. 14, pp. 537–547, Sept. 1995.

[22] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[23] M. Jordan, and R. A. Jacobs, "Hierarchical mixture of experts and EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[24] L. Hadjiiski and P. Hopke, "Design of large scale models based on multiple neural network approach," *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME, 1997, vol. 7, pp. 61–66.

[25] S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biolog. Cybern.*, vol. 23, no. 3, pp. 121–134, 1976.

[26] G. A. Carpenter and S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns," *Appl. Opt.*, vol. 26, no. 23, 1, pp. 4919–4930, Dec. 1987.

[27] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, vol. 4, no. 4, pp. 493–504, 1991.

[28] G. A. Carpenter and S. Grossberg, "Integrating symbolic and neural processing in a self-organizing architeture for pattern recognition and prediction," in *Artificial Intelligence and Neural Networks: Steps toward Principled Integration*. New York: Academic, 1994.

[29] G. A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," *Neural Networks*, vol. 11, no. 2, pp. 323–336, Mar. 1998.

[30] Y. Xie, P. K. Hopke, and D. Wienke, "Airborne particle classification with a combination of chemical composition and shape index utilizing an adaptive resonance artificial neural network," *Environ. Sci. Technol.*, vol. 28, no. 11, pp. 1921–1928, 1994.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, pp. 610–621, Nov. 1973.

[32] M. M. Galloway, "Texture analysis using gray level run length," *Comput. Graph. Image Processing*, vol. 4, pp. 172–179, 1975.

[33] M. J. Norusis, *SPSS Professional Statistics 6.1*. Chicago, IL: SPSS, 1993.

[34] M. M. Tatsuoka, "Multivariate Analysis," *Techniques for Educational and Psychological Research*. New York: Macmillan, 1988.

[35] C. E. Metz, "ROC methodology in radiographic imaging," *Invest. Radiol.*, vol. 21, pp. 720–733, 1986.

[36] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binomial ROC curve from continuously distributed test results," presented at the *1990 Annu. Meeting American Statistical Association*, Anahaim, CA, 1990.

[37] B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, and L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *Proc. SPIE*, vol. 3661, pp. 499–510, 1999.

Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers

Heang-Ping Chan[a)] and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

Robert F. Wagner
Center for Devices and Radiology Health, FDA, Rockville, Maryland 20852

Nicholas Petrick
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

(Received 14 June 1999; accepted for publication 16 September 1999)

Classifier design is one of the key steps in the development of computer-aided diagnosis (CAD) algorithms. A classifier is designed with case samples drawn from the patient population. Generally, the sample size available for classifier design is limited, which introduces variance and bias into the performance of the trained classifier, relative to that obtained with an infinite sample size. For CAD applications, a commonly used performance index for a classifier is the area, A_z, under the receiver operating characteristic (ROC) curve. We have conducted a computer simulation study to investigate the dependence of the mean performance, in terms of A_z, on design sample size for a linear discriminant and two nonlinear classifiers, the quadratic discriminant and the backpropagation neural network (ANN). The performances of the classifiers were compared for four types of class distributions that have specific properties: multivariate normal distributions with equal covariance matrices and unequal means, unequal covariance matrices and unequal means, and unequal covariance matrices and equal means, and a feature space where the two classes were uniformly distributed in disjoint checkerboard regions. We evaluated the performances of the classifiers in feature spaces of dimensionality ranging from 3 to 15, and design sample sizes from 20 to 800 per class. The dependence of the resubstitution and hold-out performance on design (training) sample size (N_t) was investigated. For multivariate normal class distributions with equal covariance matrices, the linear discriminant is the optimal classifier. It was found that its A_z-versus-$1/N_t$ curves can be closely approximated by linear dependences over the range of sample sizes studied. In the feature spaces with unequal covariance matrices where the quadratic discriminant is optimal, the linear discriminant is inferior to the quadratic discriminant or the ANN when the design sample size is large. However, when the design sample is small, a relatively simple classifier, such as the linear discriminant or an ANN with very few hidden nodes, may be preferred because performance bias increases with the complexity of the classifier. In the regime where the classifier performance is dominated by the $1/N_t$ term, the performance in the limit of infinite sample size can be estimated as the intercept ($1/N_t=0$) of a linear regression of A_z versus $1/N_t$. The understanding of the performance of the classifiers under the constraint of a finite design sample size is expected to facilitate the selection of a proper classifier for a given classification task and the design of an efficient resampling scheme.
[S0094-2405(99)00212-6]

Key words: computer-aided diagnosis, classifier design, linear classifier, quadratic classifier, neural network, sample size, feature space dimensionality, ROC analysis

I. INTRODUCTION

With the advent of digital imaging modalities, computer-aided diagnosis (CAD) is becoming an important area of research in medical imaging. A CAD algorithm can detect abnormalities and classify disease or normal cases based on image and/or patient information, and thus provide a second opinion to the radiologist in the detection or diagnostic decision making process.

Design of classifiers that can accurately distinguish normal and abnormal features is a critical step in the development of CAD algorithms. It has been shown that the performance of a classifier for unknown cases depends on the sample size used for traiñing.[1] When a finite design (training) sample size is used, the performance is pessimistically biased in comparison to that obtained from an infinitely large design sample. In order to design a classifier with a performance generalizable to the population at large, one has to use a sufficient number of case samples that are representative of the population. However, the availability of case samples is often limited in medical imaging research. It is therefore important to study the sample-size dependence of different classifiers and determine the most efficient way of training a classifier, under the constraint of a finite sample size.

We note that the concept of generalizability may be used in several technical senses when assessing the performance of a classifier: one with respect to mean classifier performance, the other with respect to the variance of classifier performance. In many classifier design problems, one is most interested in investigating if the mean performance of a classifier estimated from a given set of finite design samples can be generalized to classification performance with unknown test samples drawn from the same population of cases. The generalizability in this regard can be observed from the biases of the mean performances in the finite design set and in the test set in comparison to the optimal performance estimated from an infinite design set. The bias in the mean performance of different classifiers under various input conditions is the subject of investigation in this study. We will discuss further other interpretation of generalizability in the Discussion section of this paper.

A number of investigators have studied the finite-sample-size problem[1–9] Fukunaga[1,3] derived a general formulation for the bias and variance of a function, *f*, which is to be estimated from the available samples. When *f* is a nonlinear function of the mean vectors and covariance matrices of two feature distributions, it has been shown that a bias results from the nonlinear propagation of the finite-sample variances in the estimates of the mean vectors and covariance matrices of the distributions through this function. For multivariate-normal data, these variances are proportional to $1/N_t$, where N_t is the design sample size, and this dependence propagates into the lowest-order terms in the bias. The bias is independent of the test sample size, N_{test}. All measures of classifier performance that count the fraction of times the decision value for an abnormal case exceeds that for a normal case (independent of underlying distribution), and various measures of error for normally distributed decision functions, are nonlinear functions of the parameters of the underlying distributions. They are thus subject to this effect. Fukunaga and Hayes[3] analyzed the finite sample effects on the probability of misclassification (PMC) of a classifier and suggested a technique that makes use of the linear dependence of PMC on $1/N_t$ to estimate the performance at $N_t \rightarrow \infty$ with a finite sample set.

For the evaluation of medical diagnostic systems, the most commonly used performance index is the area under the receiver operating characteristic (ROC) curve, A_z. We have derived analytically that, for linear discriminant classifiers, the classifier performance in terms of A_z can be approximated by a linear function in $1/N_t$, under conditions when higher order terms in N_t can be neglected. We have been investigating the dependence of A_z on sample size by simulation studies.[7–9] Wagner *et al.*[10,11] have also analyzed the effects of design and test sample sizes on the variance components of the classifier performance. Although these behaviors depend strongly on the class distributions and the properties of the classifier, the studies will provide some insight into the sample size requirements for the design of different classifiers. This work may eventually lead to the selection of an efficient resampling scheme for classifier design, as well as the development of a statistical test of the sample size requirements and the generalizability of the trained classifier.

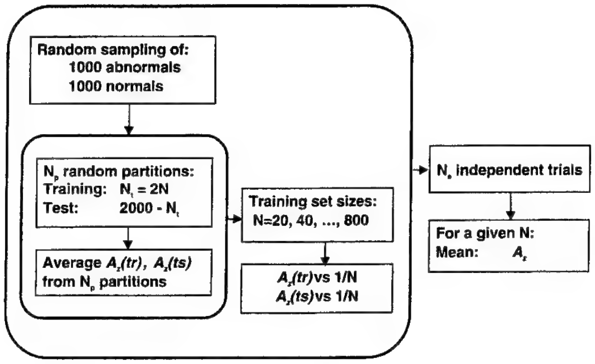


FIG. 1. The sampling and evaluation scheme of the simulation study.

In this paper, we will describe the simulation studies and analyze the effects of sample size on classifier performance. Several commonly used classifiers, including the linear discriminant, the quadratic discriminant, and the backpropagation neural network will be studied and compared under different input conditions. Feature distributions with markedly different characteristics will be used to represent a variety of situations that may be encountered in classification problems for many detection or diagnostic tasks.

II. MATERIALS AND METHODS

We performed simulation studies to evaluate the effects of sample size on classifier design. Normal and abnormal case samples were randomly drawn from known probability distributions of the two classes. These samples were then used to design classifiers for differentiation of normal and abnormal cases. The simulation approach assures that any number of case samples can be obtained from populations with known statistical properties. It thus allows evaluation of the dependence of classifier performance on design sample size and comparison of the performance with theoretically predicted optimal classification based on the chosen probability distributions.

A. Simulation study

The sampling and evaluation scheme of the simulation study is shown in Fig. 1. In this study, we considered only the situation in which equal numbers $(=N_{\text{total}}/2)$ of normal and abnormal cases randomly drawn from the class distributions were available in our data set. A resampling strategy similar to the technique suggested by Fukunaga and Hayes was devised to generate the A_z-vs-$1/N_t$ curve. Subsets of $N_{t_1}, N_{t_2}, \ldots, N_{t_j}$ design samples were randomly drawn from the available sample set, again under the constraint that the numbers of normal and abnormal samples were equal in each subset, i.e., $N_{t_i,\text{normal}} = N_{t_i,\text{abnormal}} = N_{t_i}/2\,(i=1,\ldots,j)$. A classifier was designed by using each subset of samples. The random sampling of a given subset from the available set of N_{total} samples was performed without replacement, whereas the random sampling of different subsets always started from the same set of N_{total} samples. Therefore, after drawing a given design subset N_{t_i}, the remaining samples, $N_{\text{total}}-N_{t_i}$ were independent of the design samples and used as the test samples. For simplicity, the number of design samples per class is denoted as N in the following discussion.

In general, there are two methods, resubstitution and hold-out, for testing classifier performance. In the resubstitution method, the design sample set is resubstituted into the trained classifier to test its performance, whereas in the hold-out method, an independent test set is used. It has been shown[1] that, for a Bayes classifier, if the classifier is trained with a finite number of design samples, the resubstitution estimate of the classifier performance is optimistically biased whereas the hold-out estimate is pessimisticaly biased in comparison to that achievable with an infinite design sample set. The mean performance obtained from the former estimation provides an upper bound and that from the latter provides a lower bound on the true classifier performance. When the design sample size is limited, it is important to evaluate the hold-out performance to avoid an overly optimistic prediction of the classifier performance. In the limit of very large sample size, the upper and lower bounds converge towards the unbiased estimate.

In this study, we evaluated the performance of the classifier using both the resubstitution and the hold-out methods as a function of finite design sample size N_t. In order to reduce the variances in the estimates of A_z, we randomly resampled without replacement each N_{t_i} from the same N_{total} samples N_p times, trained and tested the classifier, and estimated the average A_z from the N_p individual A_z's as shown in Fig. 1. The resubstitution or hold-out A_z-vs-$1/N_t$ curve was plotted from the j points and the unbiased estimate of A_z in the limit of $N_t \rightarrow \infty$ could be extrapolated from either curve.

This method of estimating classifier performance at large N_t by generating a few data points at finite sample sizes is similar to the Fukunaga and Hayes technique. However, we did not assume that the j points were in the linear region of the A_z-vs-$1/N_t$ curve and we used resampling to reduce the variances. In fact, one of the goals of this study was to investigate the range of design sample size in which the performance curve was approximately linear for various classifiers and probability distributions of the class populations. Therefore, we used a much larger total number of samples ($N_{\text{total}}=2000$) in our simulation study than was generally available for classifier design. We could then choose N_{t_i} over a wide range and study the behavior of the entire A_z-vs-$1/N_t$ curve.

To estimate the population mean of A_z at each N_{t_i}, we repeated the above experiment N_e times, each with 2000 independently drawn samples from the population. The population mean of A_z was estimated by averaging the A_z values obtained from the N_e experiments. We did not analyze the variances in this study because of the complication in the correlation among the N_p values of A_z introduced by resampling. A detailed analysis of the variances and its modeling was performed in a separate study by Wagner *et al.*[10,11] in which a different study design was used.

By varying the number of design samples per class, N, over a large range from 20 to 800, the regime where the $1/N_t$ dependence dominated could be observed from the A_z (population mean)-vs-$1/N_t$ (or $1/N$) curves. It is important to note that, although the number of test samples, $N_{\text{test}_i}=2000-N_{t_i}$, varied from point to point on both the resubstitution and the hold-out curves, the bias in A_z is independent of N_{test_i}.[1] The shape of the A_z-vs-$1/N$ curve is independent of N_{test_i} after N_{t_i} is fixed. However, the variance of a given A_z does depend on the test sample size.

For simplicity, we will refer to these estimates of A_z (population mean) as $A_z(\text{tr})$ for the resubstitution and as $A_z(\text{ts})$ for the hold-out performance in the following discussions.

B. Class distributions

1. Multivariate normal distributions

For three of the four types of class distributions, we assumed that the normal and abnormal classes followed multivariate normal distributions in the feature space. The dimensionality of the feature space, k, was varied from 3 to 15. The characteristics of the multivariate normal distributions can be completely specified by the multivariate mean vector of the rth class, denoted as μ_r $(r=1,2)$ and its covariance matrix, denoted as Σ_r. The separation of the normal and abnormal classes is measured by the Bhattacharyya distance, B, defined as[1,12]

$$B=\frac{1}{8}\Delta+\frac{1}{2}\ln\frac{\det[(\Sigma_1+\Sigma_2)/2]}{\sqrt{\det\Sigma_1}\sqrt{\det\Sigma_2}}, \tag{1}$$

where $\det\Sigma_r$ denotes the determinant of Σ_r, and Δ is the squared Mahalanobis distance,[12] defined as

$$\Delta=(\mu_2-\mu_1)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_2-\mu_1). \tag{2}$$

The Mahalanobis distance is the Euclidean distance between the means of the two distributions, normalized by the square root of the average of their covariance matrices. It can therefore be considered to be a measure of the signal-to-noise ratio (SNR) between the abnormal and the normal distributions. The second term of B is the contribution from the difference in the covariance matrices of the two class distributions. If the covariance matrices are equal, the second term will be zero and the Bhattacharyya distance will be equal to 1/8 of the squared Mahalanobis distance.

In the current study, three types of multivariate normal class distributions were considered. In the following discussion, we shall refer to the use of simultaneous diagonalization for the two covariance matrices of the class distributions. This operation leaves the normal-based decision functions unchanged because the distance measures that arise in these decision functions are invariant to any non-singular linear transformation.[1]

(1) Equal covariance matrices and unequal means: In this case, the covariance matrices of the normal and abnormal class distributions can be simultaneously diagonalized

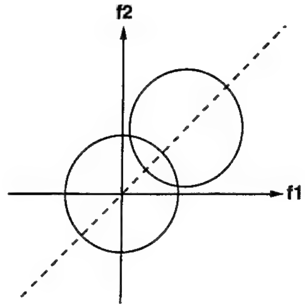


FIG. 2. A schematic illustration of the two class distributions with equal covariance matrices and unequal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

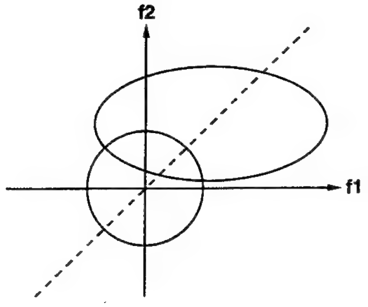


FIG. 3. A schematic illustration of the two class distributions with unequal covariance matrices and unequal means in a 2D feature space. The closed curves represent contours of equal probability in each distribution.

and the variances of the individual feature components can be scaled to unity. Therefore, without loss of generality, the covariance matrices of the two classes could be assumed to be equal to identity matrices, $\Sigma_1=\Sigma_2=I$. The mean feature vector for the first class was assumed to be zero, $\mu_1=\mathbf{0}$, and the mean feature vector for the second class, $\mu_2=\mathbf{M}$ with all components of $\mathbf{M}$ equal to a constant m. The magnitude of m could be adjusted to obtain a desired separation of the two classes. For the purpose of this simulation study, we chose m such that the squared Mahalanobis distance was 3, i.e., the Bhattacharyya distance was 3/8, for feature spaces of any dimensionality. As discussed below, this separation corresponds to a theoretical A_z of 0.89, which is in the performance range of many classification problems in CAD applications. An example of the two class distributions in a 2D feature space is shown schematically in Fig. 2.

(2) Unequal covariance matrices and unequal means: The covariance matrix of the first class was again diagonalized and scaled to be an identity matrix, $\Sigma_1=I$, and the mean feature vector for the first class was assumed to be zero, $\mu_1=\mathbf{0}$. The covariance matrix of the second class, Σ_2, was simultaneously diagonalized to have eigenvalues λ_i, $i=1,...,k$. For this study, we generated the values of λ_i with the simple relationship:

$$\lambda_i=\lambda_{\min}+\frac{(i-1)(\lambda_{\max}-\lambda_{\min})}{(k-1)},\quad i=1,...,k \tag{3}$$

and evaluated one condition where $\lambda_{\min}=1$, and $\lambda_{\max}=2$ for all dimensionalities of the feature spaces. We also assumed that the components of the mean feature vector μ_2 were equal, the values of which were adjusted to achieve a Bhattacharyya distance of 3/8. For the purpose of demonstrating the general trends of the A_z-vs-$1/N$ curves and comparing the relative performance of the different classifiers under the various conditions, the specific choices of these values are not critical. Figure 3 illustrates an example of the two class distributions in a 2D feature space.

(3) Unequal covariance matrices and equal means: The covariance matrix of the first class was the same as that in the first two cases described above. The covariance matrix of the second class was proportional to the identity matrix, $\Sigma_2=\alpha I$, where the proportionality constant α was adjusted to provide a Bhattacharyya distance of 3/8. The mean feature vectors of the two classes were equal, $\mu_1=\mu_2=\mathbf{0}$. In this case, the discriminatory power of the two classes comes entirely from the difference in the covariance matrices. A schematic of the two class distributions in a 2D feature space is shown in Fig. 4.

2. Checkerboard distributions

The fourth type of class distributions was a checkerboard where the normal and abnormal classes were located in alternate square box regions of the feature space. Within each box of the checkerboard, the feature vectors were uniformly distributed. The two classes did not overlap with each other so that they could be perfectly separated by an ''ideal'' classifier with $A_z=1$. We considered a 2×3 checkerboard in a 2D feature space and a 2×2×2 checkerboard in a 3D feature space. The example of a 2×3 checkerboard in a 2D feature space is shown in Fig. 5. Such class distributions may not be common in actual classification problems encountered in CAD. However, it was included in this study to demonstrate the capability and limitations of the different classifiers when the class distributions were not multivariate normal.

C. Classifiers

We studied three types of classifiers: the linear discriminants, the quadratic discriminants, and the back-propagation neural networks. They represent a range of classifiers commonly used in the field of pattern recognition at present.

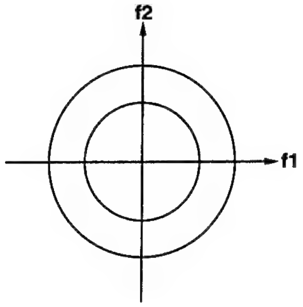


FIG. 4. A schematic illustration of the two class distributions with unequal covariance matrices and equal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

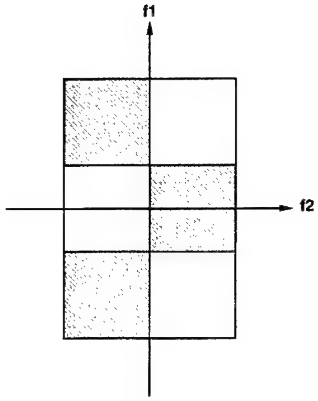


FIG. 5. An example of a 2×3 checkerboard in a 2D feature space.

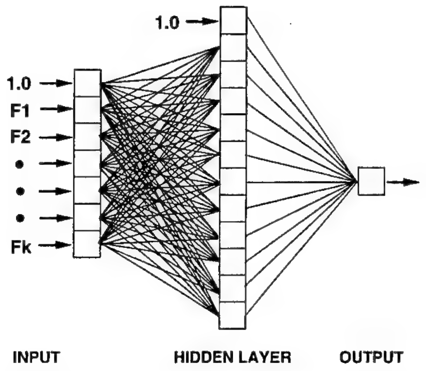


FIG. 6. A schematic diagram of a backpropagation neural network with one hidden layer.

(1) Linear discriminant classifier: The linear discriminant classifier can be derived from the means and the covariance matrices of the class distributions as follows:[1,13]

$$h_l(X)=(\mu_2-\mu_1)^T\bar{\Sigma}^{-1}X+\tfrac{1}{2}(\mu_1^T\bar{\Sigma}^{-1}\mu_1-\mu_2^T\bar{\Sigma}^{-1}\mu_2), \quad (4)$$

where $\bar{\Sigma}=(\Sigma_1+\Sigma_2)/2$, and X is the feature vector to be classified. The means and covariance matrices have to be estimated as the sample means and sample covariance matrices from the available design samples. The sample means and covariance matrices undergo a nonlinear transformation to become the discriminant scores, which in turn are transformed nonlinearly into a measure of the performance. The variances in the estimated parameters propagate into the mean classifier performance and result in a bias through the second derivative of the transformation function.

It is known that, for multivariate normal distributions with equal covariance matrices, the linear discriminant classifier is optimal and the classifier performance in the limit of large design samples is determined by the Mahalanobis distance, given by

$$A_Z=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{\Delta/2}}e^{-u^2/2}du. \quad (5)$$

For the class distributions with $\Delta=3$ to be used in this study, it can be derived from Eq. (5) that the maximum A_z that the optimal linear discriminant can achieve in the limit of large design samples is 0.89.

(2) Quadratic discriminant classifier: The quadratic discriminant classifier can be expressed as[1]

$$h_q(X)=\frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)$$
$$-\frac{1}{2}(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)+\frac{1}{2}\ln\frac{\det\Sigma_1}{\det\Sigma_2}. \quad (6)$$

When the class distributions are multivariate normal with unequal covariance matrices, the quadratic discriminant classifier is optimal in the limit of large training samples. The Bhattacharyya distance gives an upper bound on the Bayes error.[1] The general properties of the linear and quadratic classifiers have been described in the literature (for example, Fukunaga[1]).

(3) Back-propagation neural network: Many different architectures and training methods have been developed for artificial neural networks (ANN)[14] in various applications. In this study, we considered only a three-layered neural network trained with a feed-forward back-propagation method. The neural network has k input nodes, n hidden nodes, one output node, and a bias node in both the input and the hidden layers. The ANN architecture is denoted as $k-n-1$. The nodes in the ANN are fully connected and are trained with a minimum sum-of-squares-error criterion. The number of weights to be estimated is equal to $n(k+1)+(n+1)$. A schematic diagram of an ANN is shown in Fig. 6.

III. RESULTS

In our simulation study, we compared the performance of the linear, quadratic, and backpropagation neural network classifiers for the different class distributions in the feature spaces of dimensionality ranging from 3 to 15. The number of repeated experiments N_e was chosen to be 20 for all cases in the multivariate normal feature spaces and 100 in the checkerboard feature space. The number of data set partitionings N_p in each experiment ranged from 1 to 20. These choices are a compromise between computation time and estimation accuracy, especially for ANN classifiers with a large number of hidden nodes in high dimensional feature spaces. As shown in the graphs discussed below, some of the performance curves may exhibit fluctuations that could be reduced by a larger number of experiments. However, the general trend of the performance curves should not be changed by the statistical uncertainties.

(1) Multivariate normal distributions—Equal covariance matrices and unequal means: For class distributions with equal covariance matrices, the linear discriminant is theoretically the optimal classifier when the design sample size is large. However, when the design sample size is small, the performances of all classifiers are biased. Figures 7(a)–7(c) show the dependence of the A_z obtained from resubstitution (training), A_z(tr), and the A_z obtained from the hold-out method (testing), A_z(ts), on 1/N for the linear, ANN, and

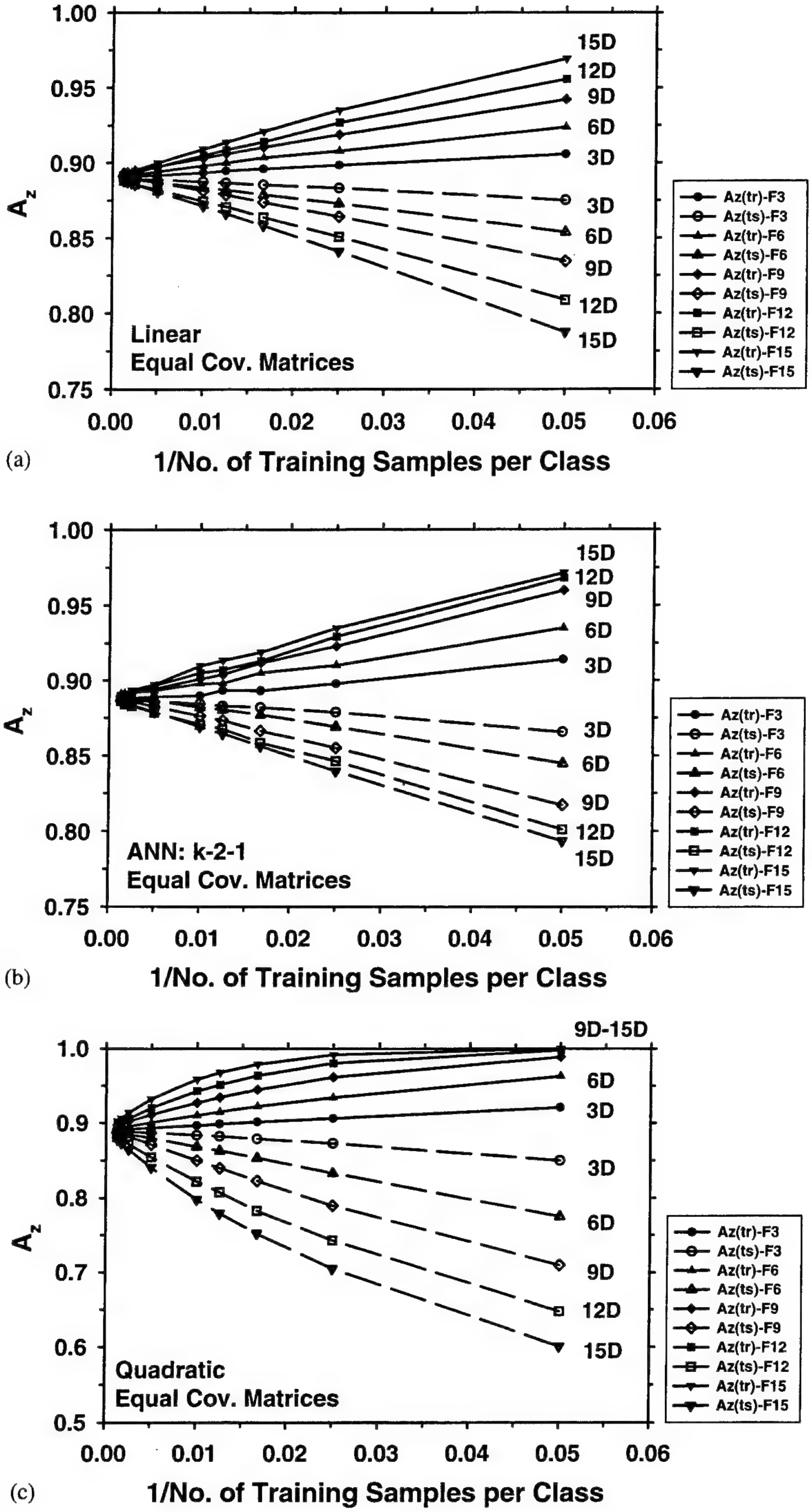


FIG. 7. The dependence of the A_z obtained from resubstitution (training-solid lines), A_z(tr), and the A_z obtained from the hold-out method (testing—dashed lines), A_z(ts), on $1/N$ for the class distributions with equal covariance matrices and unequal means. (a) Linear, (b) ANN, and (c) quadratic classifier. Legend: F3=3D feature space, etc.

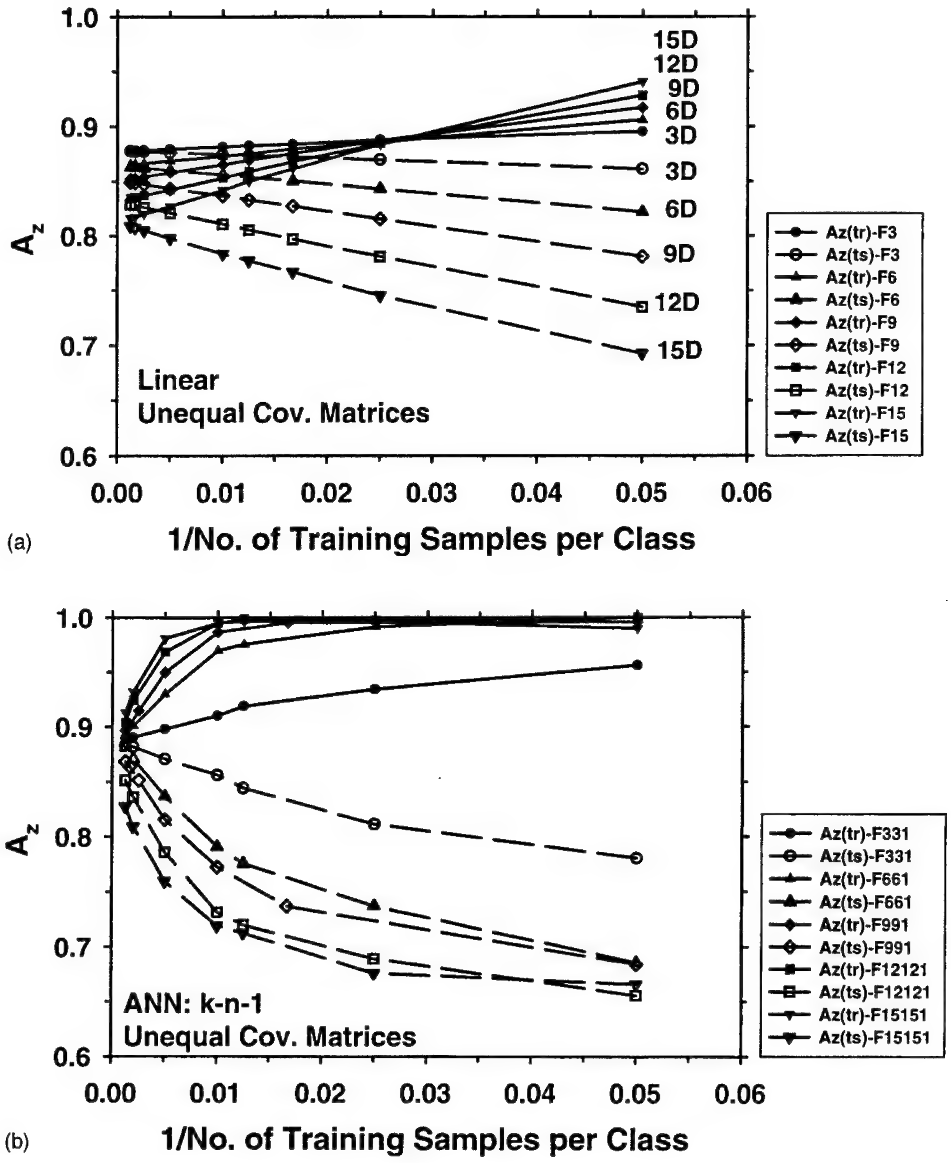


FIG. 8. The performances of the classifiers for class distributions with unequal covariance matrices and unequal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc., solid lines $=A_z(\mathrm{tr})$, dashed lines$=A_z(\mathrm{ts})$.

quadratic classifier, respectively. Two hidden nodes were used for the ANN $(k-2-1)$ because it is the smallest number of hidden nodes in a nonlinear ANN. An ANN with only one hidden node will be a linear classifier and behave in a similar manner as the linear discriminant. On the other hand, ANNs with a large number of hidden nodes (not shown) will overfit the design samples and have poor generalizability to the unknown cases, similar to the ANN curves to be discussed below. All three classifiers can reach the optimal classification accuracy of $A_z=0.89$ in the limit of large N. The curves for the linear classifier and the ANN $(k-2-1)$ at 400 training epochs (iterations) are approximately linear over the entire range. The quadratic classifier does not reach the approximately linear region until N is greater than about 100 $(1/N<0.01)$ in the higher-dimensional feature space. The biases on both the resubstitution and hold-out curves for the quadratic classifier are greater than those for the linear classifier and the ANN $(k-2-1)$. The large biases again indicate overfitting and poor generalization by the quadratic classifier in the equal-covariance-matrices situation.

(2) Multivariate normal distributions—Unequal covariance matrices and unequal means: The performances of the classifiers for class distributions with unequal covariance matrices are shown in Figs. 8(a)–8(b). The linear discriminant and the ANN $(k-2-1)$ classifier (not shown) are again approximately linear over the entire range of N studied. However, the A_z at $1/N=0$ decreases as the dimensionality of the feature space increases. This is because both the linear discriminant and the near-linear ANN $(k-2-1)$ cannot make use of the class separability due to the differences in the covariance matrices which is the second term in the Bhattacharyya distance. The second term increases relative to the first term, the squared Mahalanobis distance, when the Bhattacharyya distance is fixed and the dimensionality of the feature space increases.

The performance curves of the ANN at large N improve when a greater number of hidden nodes and a sufficient number of training epochs are used. The number of hidden nodes required to reach the optimal classification of $A_z=0.89$ at $1/N=0$ increases with the dimensionality of the feature

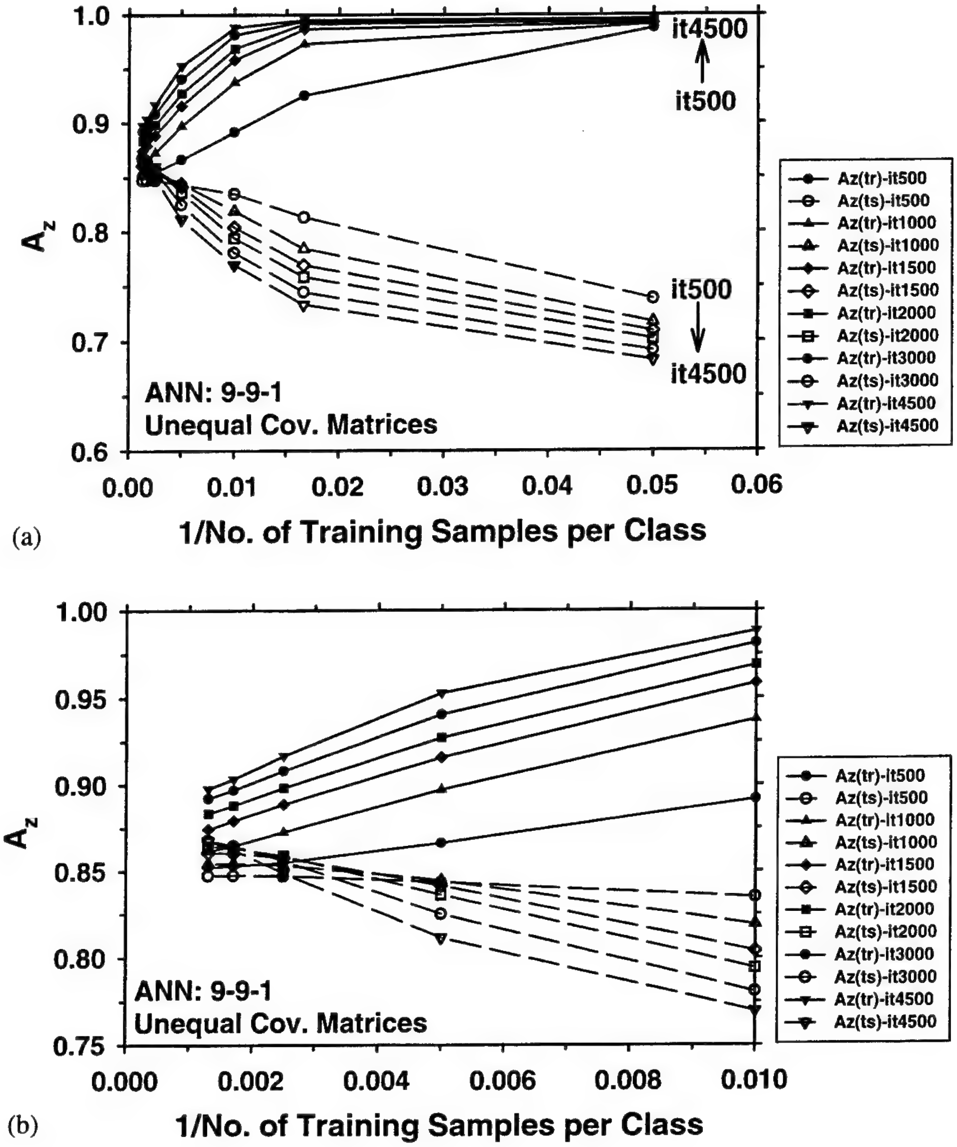


FIG. 9. The dependence of the performance curves on the number of training epochs for an ANN with nine hidden nodes in a 9D feature space: ANN(9−9−1). Legend: it500=500 training epochs, etc., solid lines=A_z(tr), dashed lines =A_z(ts). The expanded view in (b) shows the trend of the curves at large sample sizes.

space. Figure 8(b) shows the performance of the ANNs when the number of hidden nodes is equal to the dimensionality in each feature space. Since the number of weights to be trained increases rapidly with increasing number of nodes in an ANN, the number of epochs required for training the ANN to achieve a reasonable classification accuracy increases accordingly. The resubstitution and hold-out performance curves of each ANN shown in Fig. 8(b) were chosen at the smallest number of training epoch that resulted in approximately the highest A_z value when the hold-out curve was extrapolated to $1/N=0$. The number of training epochs required to reach the highest A_z increased as the dimensionality and the number of hidden nodes in the ANN increased. It ranged from about 4000 to 10 000 for the conditions shown in Fig. 8(b). We did not attempt to perform an exhaustive search for the "optimal" number of hidden nodes in each feature space because of the extensive computation time required for the search. Instead, we evaluated ANNs with a few different numbers of hidden nodes in each feature space and chose the "best" ANN within those studied. With this approximation we observed that, in a *k*-dimensional feature space and with these class distributions, an ANN with approximately *k* hidden nodes can approach the optimal performance when the design sample size and the number of training epochs are sufficiently large, as shown in Fig. 8(b).

To illustrate the training of an ANN with a large number of hidden nodes, we show the dependence of the resubstitution and the hold-out curves on the number of training epochs for ANN (9−9−1) in Fig. 9. A number of commonly discussed problems of an ANN can be observed. In the small N region below about 60 samples per class, over-parametrization and over-training are obvious, i.e., near perfect classification during training [A_z(tr) greater than 0.95] and poor generalization [A_z(ts) below about 0.8]. The problem becomes more pronounced with an increasing number of training epochs. In the middle range of 200 to 400 samples per class where A_z(ts) increases to a maximum then decreases with further training, an "optimal" number of training epoch exists. Only in the region with a sufficiently large N (greater than about 500 per class), A_z(ts) increases with

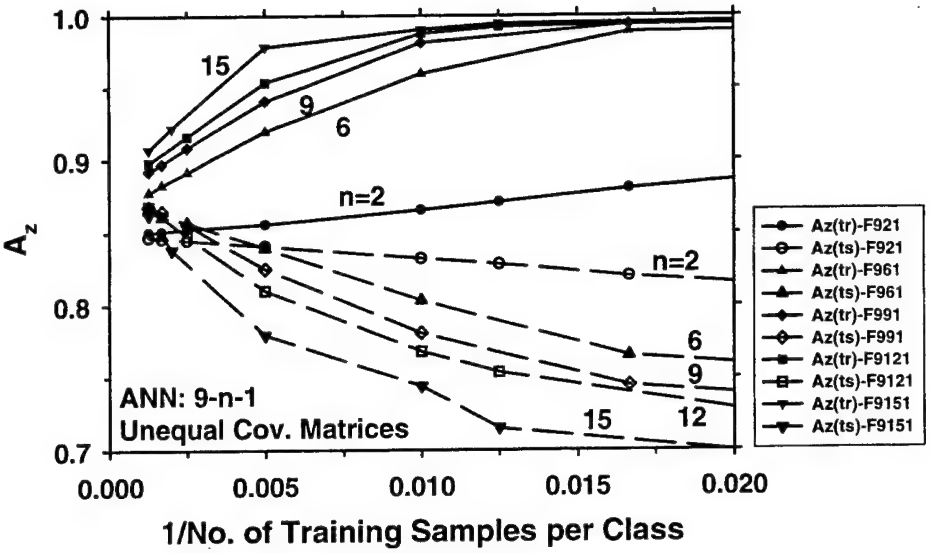


FIG. 10. The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legend: F921=ANN with two hidden nodes, etc., solid lines=A_z(tr), dashed lines =A_z(ts).

increasing number of training epochs within the range studied. The A_z(ts)-vs-$1/N$ curve becomes linear for N greater than about 200. This dependence of ANN on training epoch is generally observed for ANNs with a large number of hidden nodes and in high-dimensional feature spaces, although the design sample size required in order to avoid overtraining and over-parametrization varies. It reinforces our general experience that the ANNs with a large number of weights can overfit the design samples easily and provide poor generalization when the sample size is small.

The performance curves of ANNs with different numbers of hidden nodes in the 9D feature space are shown in Fig. 10. The curves for a given ANN were again chosen at a training epoch in which the hold-out curve approached approximately the highest performance at $1/N=0$. The chosen training epoch ranged from 600 to 12 000 for the 2- to 15-hidden-node ANNs shown. When the number of hidden nodes is small, the highest A_z obtained by extrapolation to $1/N=0$ appears to be below the theoretical optimum of 0.89. For example, the A_z extrapolated to $1/N=0$ is about 0.85 for ANN $(9-2-1)$, and is about 0.87 for ANN $(9-6-1)$. The ANN with nine hidden nodes appears to approach the optimal A_z of 0.89 in the limit of $1/N=0$. However, the ANN $(9-9-1)$ does not reach the approximately linear region until N is greater than about 200 (easier to see in Fig. 9). As can be seen from the hold-out curves, increasing the number of hidden nodes further will increase overfitting, reduce generalizability, and increase train time without gaining true improvement in performance for classification of unknown case samples.

The quadratic classifier is the theoretically optimal classifier for the class distributions with unequal covariance matrices. It can optimally utilize the class separability contributed by both the differences in the means and the covariance matrices. The performance curves for the quadratic classifier (not shown) in feature spaces of different dimensionalities are very similar to those obtained for the equal covariance matrices situation [Fig. 7(c)]. The A_z of the quadratic classi-

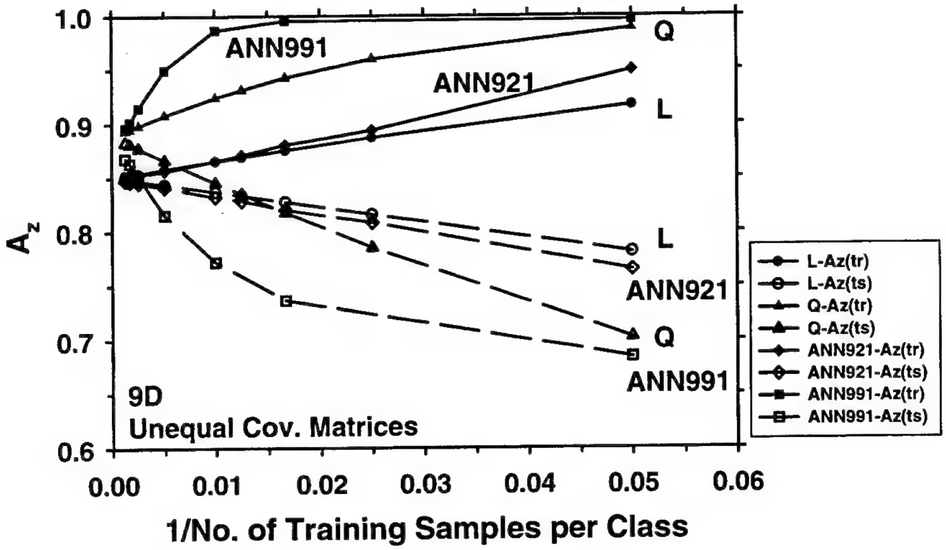


FIG. 11. Comparison of the performance curves of the linear, quadratic, ANN(9−2−1), and ANN(9−9−1) classifiers in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legends: L=linear; Q=quadratic, ANN=neural network, solid lines=A_z(tr), dashed lines=A_z(ts).

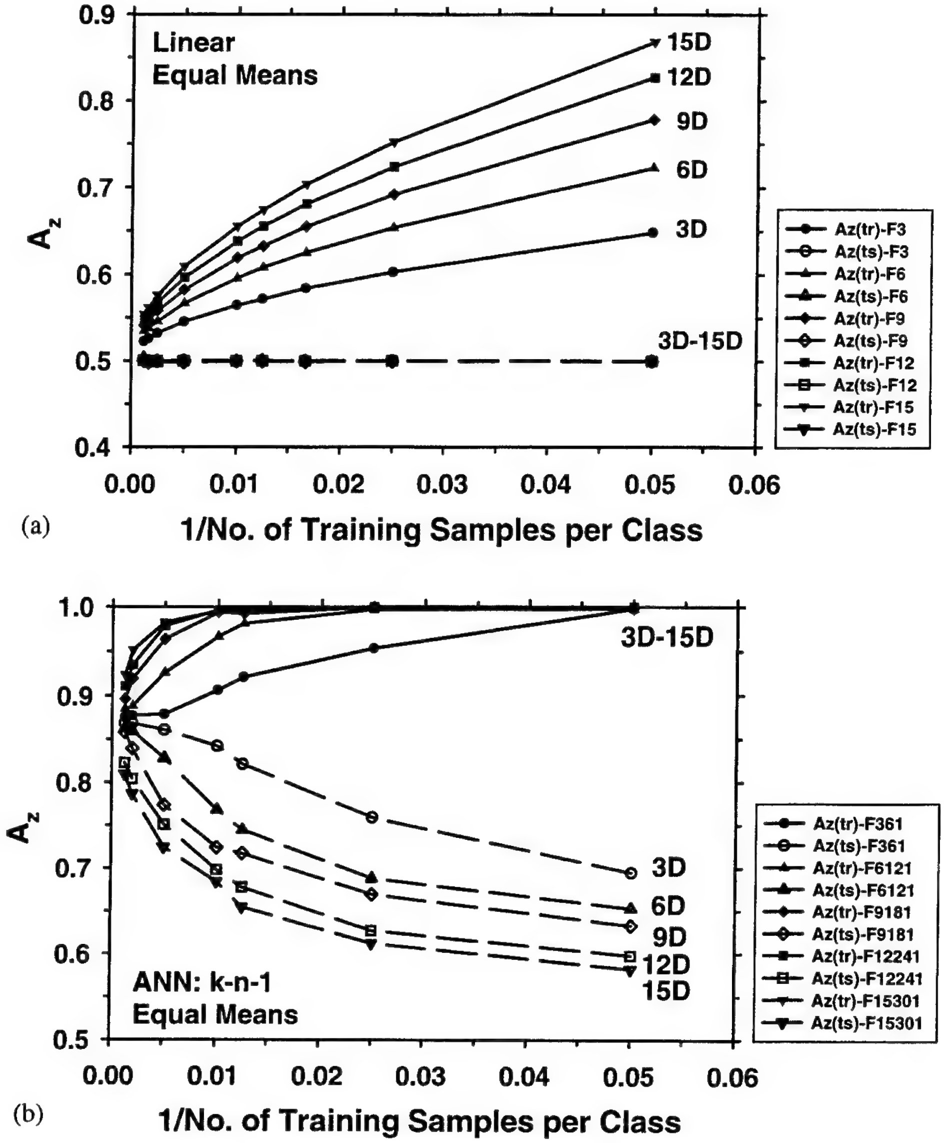


FIG. 12. The dependence of the performance curves on dimensionality of feature space for the class distributions with unequal covariance matrices and equal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc. F921 =ANN with two hidden nodes, etc. solid lines=A_z(tr), dashed lines=A_z(ts).

fier reaches the optimal value of 0.89 in the limit of large N for all dimensionalities studied.

Figure 11 shows a comparison of the performance of the linear, quadratic, and the ANN classifiers with two and nine hidden nodes. The biases on the resubstitution and the hold-out curves of the quadratic classifier are not as large as those of the ANN $(9-9-1)$ classifier. However, in the regime of small design sample sizes, the hold-out curve of the optimal quadratic classifier can be much lower than the corresponding curves of the linear classifier or ANN with one or two hidden nodes. This result indicates that the theoretically optimal classifier may not be the optimal choice when the available design sample size is small and over-parametrization becomes an important consideration.

(3) Multivariate normal distributions—Unequal covariance matrices and equal means: Figure 12(a) shows the dependence of A_z on $1/N$ for the linear classifiers for the class distributions with equal means. Since the Mahalanobis distance is zero when the means of the two class distributions are equal, the linear classifier performs no better than random guessing in the hold-out situation ($A_z(\mathrm{ts})=0.5$). However, it is somewhat surprising that the resubstitution curve can be biased to very high A_z values, when the design sample is small. The bias increases with increasing dimensionality of the feature space because the severity of overfitting to the design samples worsens with increased parameterization in the linear discriminant function. This indicates that the predicted performance of a classifier can be unrealistically optimistic if the test samples are not independent of the design samples.

For the class distributions with equal means, it is much more difficult to train the ANN classifier. The number of hidden nodes and the number of training epochs required for the ANN to approximate the decision surfaces, which are spherical hypersurfaces in the k-dimensional feature space, increase as k increases. Figure 12(b) shows the A_z-vs-$1/N$ curves for the ANNs in which the number of hidden nodes is 2 times the dimensionality of the feature space. The number of training epochs required to approach the highest perfor-

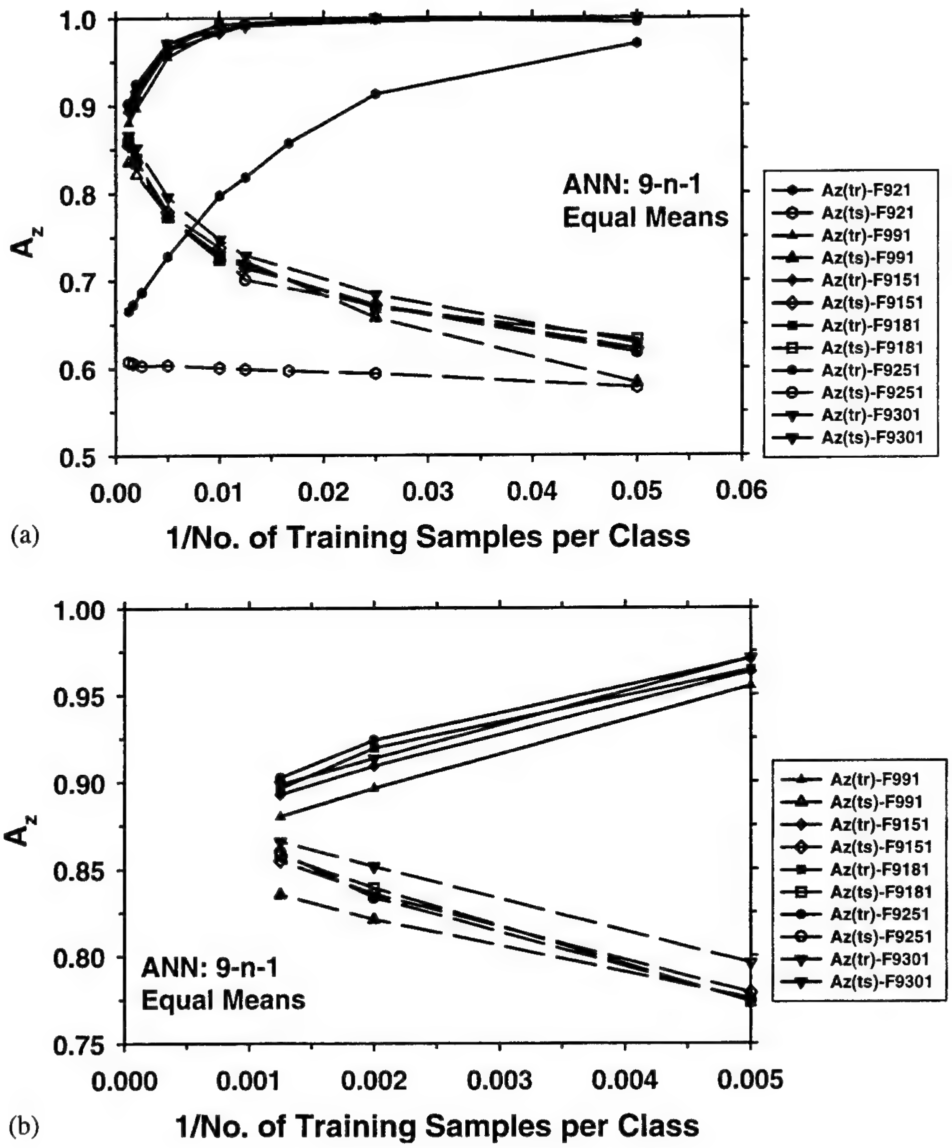


FIG. 13. (a) The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and equal means. In the expanded scale (b), the approximately linear regions of the curves can be observed. Solid lines $=A_z(\mathrm{tr})$, dashed lines $=A_z(\mathrm{ts})$.

mance for a given ANN architecture ranges from about 1800 to 20 000 in these cases. Again we did not attempt an exhaustive search for the "optimal" number of hidden nodes in each case. These ANNs were chosen because they appear to approach the maximum performance of $A_z=0.89$ in the limit of large N and their number of hidden nodes is a simple multiple of the dimensionality. Compared to the class distributions with unequal means, for a given dimensionality, the number of hidden nodes and the number of training epochs required for achieving the near maximum performance at large N are greater in this equal-mean situation. Figure 13(a) shows an example of the dependence of the performance curves on the number of hidden nodes in the 9D feature space. Figure 13(b) is an enlarged view of the curves in Fig. 13(a) in the range where the sample size is greater than 200 per class. The hold-out performance of ANN(9−9−1) at $1/N=0$ reaches about 0.85. When the number of hidden nodes is greater than nine, the performances of the ANNs at $1/N=0$ are similar and approach the optimal A_z.

The quadratic discriminant is again the theoretically optimal classifier for the class distributions with unequal covariance matrices. Its performance curves (not shown) are very similar to those plotted in Fig. 7(c), except that the extrapolated A_z values at $1/N=0$ do not reach as high as those in the equal covariance matrices situation. By using the approximately linear region of the A_z-vs-$1/N$ curve at N greater than 100, the extrapolated A_z ranges from about 0.873 to 0.885 for the 3D to 15D feature spaces. In this case, it is much more efficient to train a quadratic discriminant than the ANN. Since the linear discriminant and ANNs with few hidden nodes cannot provide effective classification regardless of the design sample size, the quadratic discriminant is obviously the optimal classifier both in terms of performance and training efficiency.

(4) Checkerboard distributions: In a feature space with checkerboard class distributions, classification is difficult for many classifiers because of the disjoint clusters of samples belonging to the same class. We compared the three classifiers in such a situation by two examples. Figure 14 shows the performance curves of the three classifiers in a 2D feature

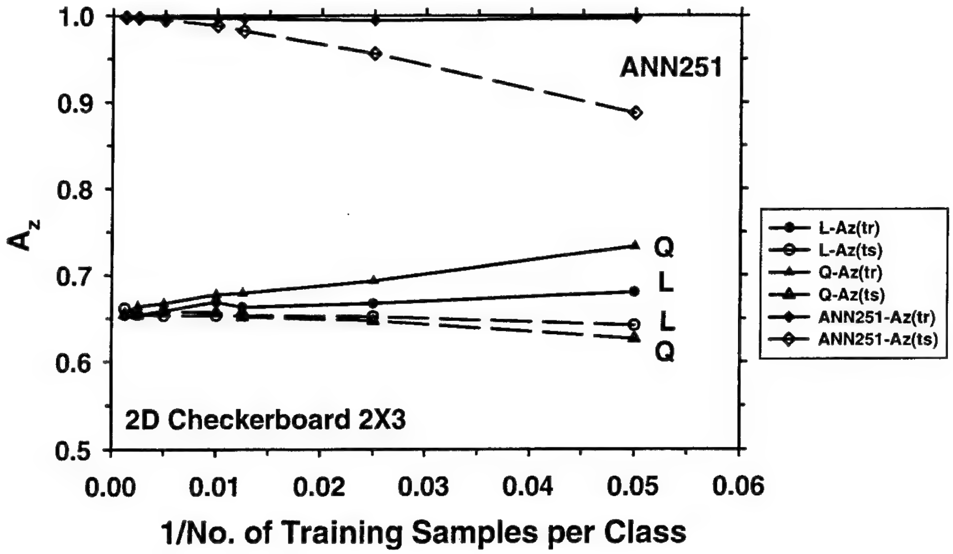


FIG. 14. Performance curves of the three classifiers for a 2×3 unit checkerboard in a 2D feature space. L=linear, Q=quadratic, ANN251=backpropagation neural network with five hidden nodes. Solid lines=A_z(tr), dashed lines=A_z(ts).

space with a 2×3 unit checkerboard distribution. Both the linear and the quadratic discriminants perform poorly even for the resubstitution method where A_z values are in the range of 0.6 to 0.7. However, the ANN(2−3−1) can achieve an A_z of 0.96 (not shown) and the ANN(2−5−1) a near-perfect classification at a training epoch of about 1200.

In a 3D feature space with a 2×2×2 unit checkerboard distribution, the difficulty in classification experienced by the linear and quadratic discriminants is even more apparent. Figure 15 shows that the hold-out curve of the linear classifier is basically the same as random guessing. The hold-out curve of the quadratic classifier is slightly higher than 0.5 at small design sample sizes but approaches 0.5 as the design sample increases. On the other hand, the ANN(3−3−1) can attain a test A_z of 0.9 (not shown) and the ANN(3−5−1) can reach near-perfect classification at large design sample sizes after about 1500 training epochs. These two examples demonstrate that an ANN classifier can be superior to the linear or quadratic classifiers for class distributions that are very different from the idealized multivariate normal distributions.

IV. DISCUSSION

Classifier design is an important field of research in computer-aided diagnosis. Yet many of the issues related to classifier design have not been explored systematically. This simulation study is a part of our on-going investigation of the sample size effects on classifier design.[7–11,15] In this study, we evaluated classifier performance for three multivariate normal class distributions with specific properties: equal covariance matrices, unequal covariance matrices, and equal means. These distributions are idealized but they do approximate a range of situations that may occur in real classification problems. Since the optimal classifier and the upper bound of classification accuracy in the limit of $1/N=0$ are

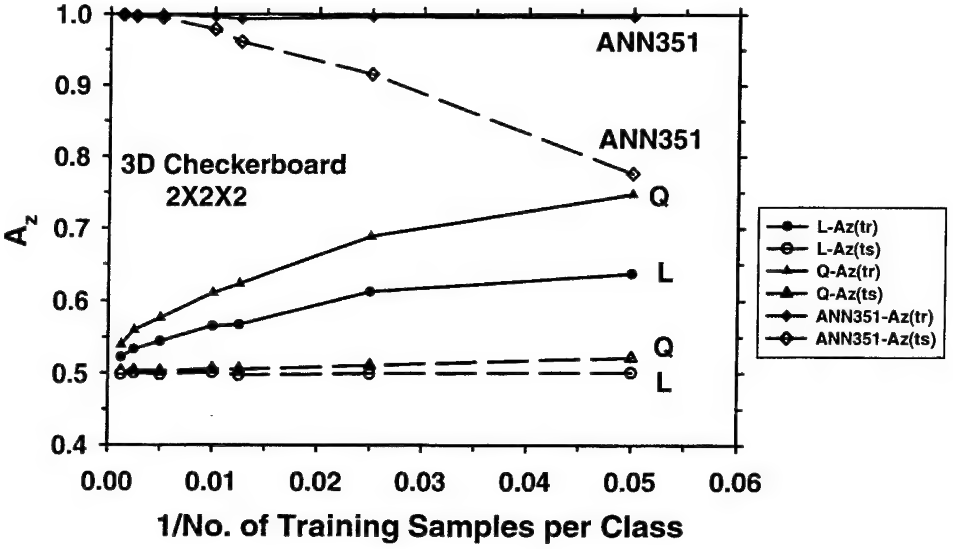


FIG. 15. Performance curves of the three classifiers for a 2×2×2 unit checkerboard distribution in a 3D feature space. Legend: L=linear, Q=quadratic, ANN351=backpropagation neural network with five hidden nodes.

known for each of these cases, we can compare the performances of the classifiers under each condition with the optimum. In addition, a checkerboard class distribution was included in the study. A comparison of the performances of the different classifiers for this class distribution can illustrate their effectiveness when the distributions are very different from multivariate normal.

For all three classifiers, the A_z(tr) obtained by resubstitution is biased optimistically while the A_z(ts) obtained by testing with an independent test set is biased pessimistically, relative to the A_z in the limit of $N\rightarrow\infty$, except for the situations when A_z(tr) is bounded from above by perfect classification ($A_z=1$) or when A_z(ts) is bounded from below by random guessing ($A_z=0.5$). The magnitude of the biases increases as the design sample size decreases and as the dimensionality of the feature space increases. In the cases where a given classifier has no discriminatory power for a given class distribution, for example, the linear discriminant for the equal-mean or checker-board class distributions, or the quadratic discriminant for the 3D checker-board class distribution, the test A_z(ts) remains almost constant at 0.5, independent of the design sample size. In many cases, the A_z-vs-$1/N$ curve cannot be approximated by a straight line that extrapolates to the A_z at $1/N=0$ until the design sample sizes are very large, beyond the range of sample sizes that are generally available for CAD classifier design. To estimate the performance of a classifier at large N under the constraint of a small design sample, one may use the Fukunaga and Hayes resampling scheme[3] to derive several points along the A_z-vs-$1/N$ curves in the small sample size region. If the extrapolated resubstitution and hold-out curves do not converge to approximately the same A_z at $1/N=0$, an average of the points on the two curves which correspond to the same design sample size may be a closer estimate of A_z than either A_z(tr) or A_z(ts). It may be noted that the resubstitution and the hold-out curves are not biased symmetrically from the A_z at infinite N, the average thus obtained will only be a rough estimate. It is also not valid in cases when the classifier has no discriminatory power with A_z(ts) constant at about 0.5 or when the resubstitution curve is overly optimistic with A_z(tr) constant at about 1.

In any case, caution should be taken in estimating classifier performance by extrapolation to $1/N=0$ or by averaging the resubstitution and hold-out performance as discussed above. The estimated performance contains variances that have to be estimated using further tools. One such attempt in estimating the components of variance by a bootstrapping resampling scheme has been studied recently by Wagner *et al.*[11] These estimates reveal the amount of bias and variance in the classifier performance obtained with the finite design samples, thus allowing estimation of the sample size required to achieve a desired degree of generalizability, rather than replacing the need for a larger sample set and further studies.

With the equal-covariance-matrix class distributions, the linear discriminant is the optimal classifier as expected. The biases are low and the computation is efficient. Moreover, since the A_z-vs-$1/N$ relationship is linear over almost the entire range of design sample sizes, the classifier performance at very large N can be estimated from the small sample size performance by linear interpolation, as suggested by Fukunaga and Hayes[3] and demonstrated previously by Wagner *et al.*[9]

With the unequal-covariance-matrices and equal-mean class distributions, the linear discriminant and the backpropagation neural network with one hidden layer are inferior to the quadratic classifier when the design sample size is large. The linear discriminant cannot utilize the difference in the covariance matrices and underestimates the class separability even when an infinite number of design samples is available. The ANN needs a relatively large number of hidden nodes and a large number of training epochs in order to reach the optimal performance. Its hold-out performance and the computation efficiency are both inferior to those of the quadratic classifier. However, for the unequal-covariance-matrices and unequal-mean case and a small design sample size, the linear classifier or an ANN with very few hidden nodes, e.g., $n=2$, provides better hold-out performance than the more complex ANNs or the optimal quadratic classifiers. These results indicate that the bias on classifier performance increases with increasing complexity (loosely related to the number of parameters to be estimated) of the classifier. The linear classifier contains $(k+1)$ independent parameters and the quadratic classifier contains $(k+1)(k+2)/2$ independent parameters in their formulations. The number of weights to be estimated for the ANN depends on the number of hidden nodes as $n(k+1)+(n+1)$. The number of weights in an ANN can therefore easily exceed that of a quadratic classifier, although the estimation of the mean and covariance matrices for the linear and quadratic discriminants may contribute additional ''complexity'' to the classifier design. Two observations can be made. First, when the available sample size is small, a simple classifier will have better generalization than a more complex classifier. Second, a complex ANN or a quadratic classifier trained with an insufficient number of design samples generalizes poorly, even if it is the optimal classifier for the class distributions. It is therefore important to select an appropriate classifier by taking into consideration the design sample size.

A further problem in classifier design is that the true population distributions of the classes in the feature space are generally unknown. It was suggested that the quantile–quantile (Q–Q) plot and the chi-square plot may be used for investigating the normality of univariate and multivariate sample distributions, respectively.[16] However, it is still unknown under what criteria the chi-square plot will indicate that it is optimal to use a classifier designed under the normality assumption. For any measure of goodness-of-fit, when the sample size is small, only the most aberrant deviations from the normal distribution can be identified as a lack of fit from these plots.[16] Therefore, there is often no *a priori* knowledge to select an ''optimal'' classifier or to predict whether the observed performance is caused by the sample size, the choice of an overly complex classifier, or by an actual poor separation of the classes in the feature space. If one observes poor generalization of a trained classifier in a truly independent test set, it will be important to take into consideration all these factors and redesign the classifier.

In this study, we assumed that the best features have already been determined for the classification task. In a general classifier design problem, the best set of features usually has to be selected based on the available design samples. The feature selection step will introduce additional biases to the classifier performance. The number of features selected also has a strong influence on the classifier design, as can be seen from the dependence of the bias on the dimensionality of the feature space. The investigation of this more complex situation including both the feature selection and classifier training steps is underway.[17]

The term generalizability is nonspecific and needs to be qualified here. The present paper is concerned with the generalizability of the mean performance of classifiers to unknown test samples drawn from the same population of cases. We have shown in this paper that the mean performance of a classifier depends on the number of samples used to train the classifier, the architecture of the classifier, and—for multivariate-normal data—the means and covariances of the population distributions. Suppose in this context that a classifier is trained on a given finite number of design samples (patients). The mean performance of the classifier over independent replications with the same number of design samples is generalizable to studies characterized by the same number of design samples. In other words, the mean resubstitution or hold-out performance is an unbiased estimate for repeated sampling of independent design and test sample sets, respectively, when the same number of design samples is used. The classifier performance may not, however, be generalizable to studies characterized by a different number of design samples. In particular, when a very large and representative design sample size is used, the mean performance may be very different from the mean performance that characterizes the finite-training-sample condition. When the mean performance under the conditions of a finite design sample size is close to that expected with a very large design sample size, the finite-training sample performance is said to be generalizable to the population performance.

The term generalizability is not only used with respect to mean performance, it is also used with respect to uncertainty in performance, as reflected in estimates of error bars (standard deviations, or the corresponding variances). For example, if we think of repeating a given training and testing experiment on a classifier and if only the test samples are drawn independently on the repeated trials, then the estimated uncertainties are said to be generalizable only to a population of test samples. If, however, we think of repeating the experiment and independently drawing new training samples as well as new test samples, then the estimated uncertainties are said to be generalizable to a population of trainers and a population of testers.[17] Models for the components of variance in both paradigms are the subjects of current work in progress.[10,11] A key point of this latter work is the fact that for computer-aided diagnosis, most available software for ROC analysis only provides estimates of uncertainty that are generalizable to a population of test samples.

In this investigation, we have limited our study to only three types of classifiers: the linear discriminant, the quadratic discriminant, and the backpropagation ANNs with one hidden layer. There are, of course, many other variations of the ANN architecture and other parametric or non-parametric classifiers available for feature classification tasks. The purpose of our work is not to exhaustively evaluate all possible combinations of class distributions and classifiers. Rather, by limiting our investigation to some well-known situations, we can perform systematic analyses and gain some insights into the classifier design problems. Furthermore, we have limited our discussion here to the estimates of the mean classifier performance. Wagner *et al.*[10,11] have investigated the variances of classifier performance estimated from a finite sample set and developed models to study the relative importance of the sizes of the training and test samples. It has been demonstrated that a components-of-variance model can be estimated with a finite sample set by using a bootstrap method. More importantly, the analysis of variances can reveal the generalizability of the performance estimates to other training and test sample sets in the population. Our long term goals are to find some guidelines for designing efficient resampling schemes that can minimize the bias and variance of a trained classifier using the available samples, and to provide a quantitative design tool that can estimate the design sample size requirement for a larger ''pivotal'' study from the results of a smaller ''pilot'' study in order to achieve a desired precision in A_z and the desired generalizability.

ACKNOWLEDGMENTS

This work is supported in part by USPHS Grant No. CA 48129 and by a grant from the U.S. Army Medical Research and Materiel Command DAMD 17-96-1-6254, a Career Development Award (B.S.) DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command and a Whitaker Foundation Grant (N. P.). The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in this paper should be inferred. The authors are grateful to Charles E. Metz, Ph. D., for providing the LABROC1 programs.

[a]Author to whom correspondence should be addressed. Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, UHB1 F510B, Ann Arbor, MI 48109-0030; Phone: 734-936-4357; Fax: 734-936-7948; Electronic mail: chanhp@umich.edu

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990).

[2] S. Raudys and V. Pikelis, ''On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition,'' IEEE Trans. Pattern. Anal. Mach. Intell. **PAMI-2**, 242–252 (1980).

[3] K. Fukunaga and R. R. Hayes, ''Effects of sample size on classifier design,'' IEEE Trans. Pattern. Anal. Mach. Intell. **11**, 873–885 (1989).

[4] R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, in *Information Processing in Medical Imaging*, edited by H. H. Barrett and A. F. Gmitro (Springer-Verlag, Berlin, 1993).

[5] R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, "On combining a few diagnostic tests or features," Proc. SPIE **2167**, 503–512 (1994).

[6] D. G. Brown, A. C. Schneider, M. P. Anderson, and R. F. Wagner, "Effect of finite sample size and correlated/noisy input features on neural network pattern classification," Proc. SPIE **2167**, 180–190 (1994).

[7] H. P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: quadratic and neural network classifiers," Proc. SPIE **3034**, 1102–1113 (1997).

[8] H. P. Chan, B. Sahner, R. F. Wagner, and N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," Proc. SPIE **3338**, 845–858 (1998).

[9] R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Finite-sample effects and resampling plans: applications to linear classifiers in computer-aided diagnosis," Proc. SPIE **3034**, 467–477 (1997).

[10] R. F. Wagner, H. P. Chan, J. T. Mossoba, B. Sahiner, and N. Petrick, "Components of variance in ROC analysis of CADx Classifier performance," Proc. SPIE **3338**, 859–875 (1998).

[11] R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Components of variance in ROC analysis of CADx classifier performance. II: Applications of the bootstrap," Proc. SPIE **3661**, 523–532 (1999).

[12] D. J. Hand, *Discrimination and Classification* (Wiley, New York, 1981).

[13] P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).

[14] J. A. Freeman and D. M. Skapura, *Neural Networks-Algorithms, Applications, and Programming Techniques* (Addison-Wesley, Reading, 1991).

[15] H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis in mammography: effects of finite sample size," Med. Phys. **24**, 1034–1035 (1997).

[16] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1982).

[17] C. A. Roe and C. E. Metz, "Variance-component modeling in the analysis of receiver operating characteristic index estimates," Acad. Radiol. **4**, 587–600 (1997).

Heang-Ping Chan, PhD
Berkman Sahiner, PhD
Mark A. Helvie, MD
Nicholas Petrick, PhD
Marilyn A. Roubidoux, MD
Todd E. Wilson, MD
Dorit D. Adler, MD
Chintana Paramagul, MD
Joel S. Newman, MD
Sethumadavan Sanjay-Gopal, PhD

Improvement of Radiologists' Characterization of Mammographic Masses by Using Computer-aided Diagnosis: An ROC Study[1]

Index terms:
Breast neoplasms, 00.31, 00.32
Breast neoplasms, radiography, 00.111, 00.119
Breast radiography, 00.111, 00.119
Computers, diagnostic aid
Receiver operating characteristic curve (ROC)

Radiology 1999; 212:817–827

Abbreviations:
CAD = computer-aided diagnosis
PPV = positive predictive value
ROC = receiver operating characteristic

PURPOSE: To evaluate the effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses seen on mammograms.

MATERIALS AND METHODS: The authors previously developed an automated computer program for estimation of the relative malignancy rating of masses. In the present study, the authors conducted observer performance experiments with receiver operating characteristic (ROC) methodology to evaluate the effects of computer estimates on radiologists' confidence ratings. Six radiologists assessed biopsy-proved masses with and without CAD. Two experiments, one with a single view and the other with two views, were conducted. The classification accuracy was quantified by using the area under the ROC curve, A_z.

RESULTS: For the reading of 238 images, the A_z value for the computer classifier was 0.92. The radiologists' A_z values ranged from 0.79 to 0.92 without CAD and improved to 0.87–0.96 with CAD. For the reading of a subset of 76 paired views, the radiologists' A_z values ranged from 0.88 to 0.95 without CAD and improved to 0.93–0.97 with CAD. Improvements in the reading of the two sets of images were statistically significant (P = .022 and .007, respectively). An improved positive predictive value as a function of the false-negative fraction was predicted from the improved ROC curves.

CONCLUSION: CAD may be useful for assisting radiologists in classification of masses and thereby potentially help reduce unnecessary biopsies.

Breast cancer is the most prevalent non–skin cancer in women; 178,700 new cases are estimated to have occurred in 1998 (1). The mortality of breast cancer is the second highest among all cancer deaths in women (1). At present, there is no effective method to prevent breast cancer. The best approach to reducing the breast cancer mortality rate is early detection and treatment. Because the mammographic features of early-stage breast cancers are not very specific, the need for high detection sensitivity leads to biopsy of many low-suspicion lesions. The positive predictive values (PPVs) of mammographic signs are, therefore, often below 30% (2,3).

Computer-aided diagnosis (CAD) is considered to be one of the approaches that may improve the efficacy of mammography (4). With CAD, a computerized detection algorithm alerts a radiologist to the location of the suspicious lesions, and/or a trained computer classifier provides the radiologist with an estimate of the likelihood of malignancy of a lesion. The radiologist takes into consideration the information provided by the computer before making a decision. This "second opinion" may improve the diagnostic accuracy because it serves as a form of double reading (5). Furthermore, a computer evaluation is often more consistent and reproducible than a human decision maker (6).

Considerable research has been devoted to the development of computerized schemes for the detection and classification of mammographic abnormalities. These efforts have advanced the CAD technology such that clinical application appears to be possible in the

[1] From the Department of Radiology, University of Michigan Hospital, UH B1F510, 1500 E Medical Center Dr, Ann Arbor, MI 48109-0030. From the 1997 RSNA scientific assembly. Received August 10, 1998; revision requested September 8; revision received November 30; accepted January 21, 1999. Supported in part by United States Public Health Service grant CA 48129 and by U.S. Army Medical Research and Materiel Command grant DAMD 17-96-1-6254. B.S. supported by Career Development award DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command. N.P. supported by a grant from the Whitaker Foundation. **Address reprint requests to** H.P.C. (e-mail: *chanhp@umich.edu*).

The content of this article does not necessarily reflect the position of the funding agencies, and no official endorsement of any equipment or product of any companies mentioned in this article should be inferred.

Author contributions:
Guarantor of integrity of entire study, H.P.C.; study concepts and design, H.P.C., M.A.H., B.S., N.P.; literature research, H.P.C., M.A.H.; experimental studies, M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.; data acquisition, all authors; data analysis, H.P.C., B.S., N.P.; statistical analysis, H.P.C.; manuscript preparation, editing, and review, H.P.C., B.S., M.A.H., N.P., M.A.R., T.E.W., D.D.A., C.P., J.S.N.

Reprinted from RADIOLOGY, Vol. 210, No. 1, Pages 817-827, September 1999.

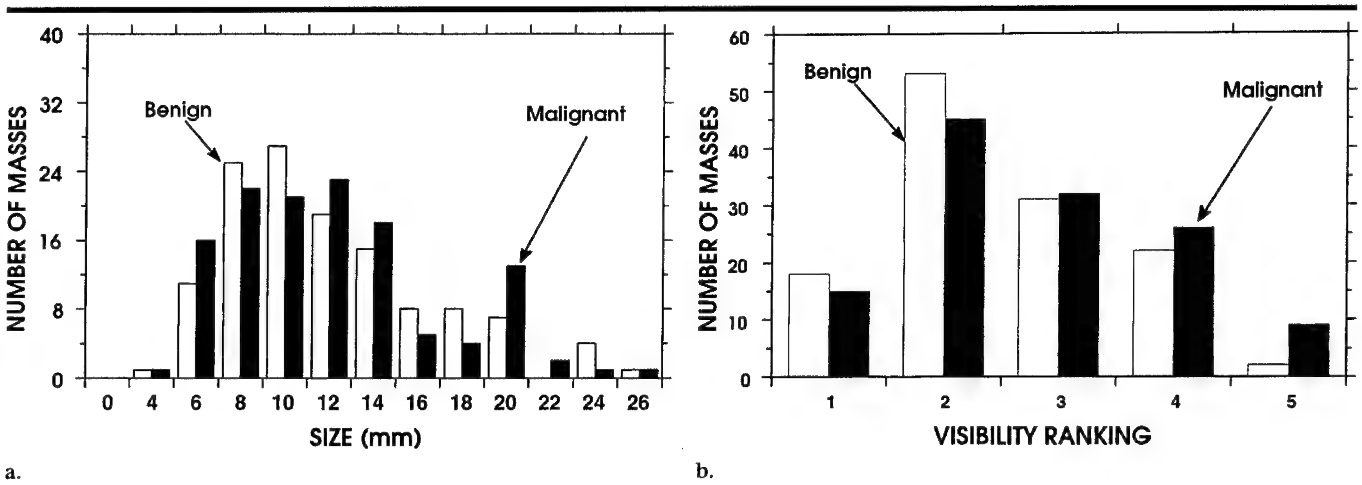


Figure 1. Histograms illustrate the distributions of (a) size (ie, length of the long axis) and (b) visibility ranking (1 = obvious, 5 = subtle) of the 253 masses included in the data set. Because classification accuracy depends on the case mix, these distributions provided some information on the masses in the data set.

near future. It is, therefore, necessary to evaluate the effects of CAD on radiologists' detection and diagnosis of mammographic lesions. In a previous receiver operating characteristic (ROC) study, we demonstrated that CAD could improve radiologists' accuracy in the detection of subtle microcalcifications on mammograms (7). Kegelmeyer et al (8) also reported an improvement in radiologists' sensitivity for the detection of spiculated masses with use of a computer aid. For the classification of mammographic lesions, it has been shown that a computer classifier that estimated the likelihood of malignancy on the basis of mammographic features extracted by radiologists could improve radiologists' accuracy in distinguishing malignant from benign lesions (9–11).

We previously conducted ROC studies to compare the performance of radiologists with that of the computer (12) and to compare radiologists' ability to classify masses with and without CAD (13). Jiang et al (14) also performed an ROC study of the effect of CAD on radiologists' performance in classifying microcalcifications. The results of all of these observer performance studies indicate the potential to improve mammographic interpretation with a computer aid.

We have developed an automated method to analyze masses seen on mammograms (15–17). A mass is segmented from its surrounding breast tissue, and an image transformation technique is used to transform the mass margin from the polar coordinate system to the Cartesian coordinate system. A linear discriminant classifier then extracts the useful texture features from the transformed image and merges them into a relative malignancy rating. Our approach is different from others that use a trained classifier to merge radiologist-extracted image features or feature codes by using the American College of Radiology Breast Imaging Reporting and Database System lexicon (9–11). Our fully automated method has the advantage that, unlike a human reader, it does not have variability in feature recognition and coding. In addition, the computer may be able to extract some information, such as texture features, that may not be readily perceived by human eyes. We conducted an ROC study to evaluate whether this computer aid can improve radiologists' performance in the classification of mammographic masses (13). The results of our observer performance study are described in this article.

Other investigators also have reported on automated algorithms for the classification of mammographic masses (18–21). The methods used in these algorithms varied, and their accuracy in classification cannot be compared directly because of the differences in the data sets. However, the effects of CAD on radiologists' performance are not expected to depend strongly on the specific algorithm if different computer aids of comparable accuracy are used. Therefore, the applications of the findings of this study should not be limited to our computerized classification aid.

MATERIALS AND METHODS

Data Set

The data set for this study consisted of 253 mammograms obtained in 103 patients. Each image contained a biopsy-proved mass that was evaluated in this study. Some cases involved multiple views or images from multiple examinations. The cases were randomly selected from patient files from the breast imaging division of a National Cancer Institute–designated national cancer center with the approval of the Institutional Review Board. The PPV of masses recommended for biopsy at this center is about 25%–30%, but an approximately equal number of malignant and benign masses (127 and 126, respectively) were chosen to enhance the statistical power in this observer performance study. Any images that were judged to be technically poor were excluded.

The mammograms were acquired with a contact technique. The dedicated mammographic systems had a molybdenum anode and molybdenum filter, a 0.3-mm nominal focal spot, and a reciprocating grid. MinR/MinR-E screen-film systems (Eastman-Kodak, Rochester, NY) were used with these units. Sixty-two of the malignant masses and six of the benign masses were judged to be spiculated by a radiologist (M.A.H.) experienced in mammography. The radiologist also measured the size (ie, longest dimension) and ranked the visibility of the masses on a scale of 1 (obvious) to 5 (subtle) relative to the range of visibility of masses encountered in clinical practice. For a description of the masses included in the data set, histograms of the size and visibility of the masses are shown in Figures 1a and 1b, respectively.

For the computer analysis, the selected mammograms were digitized with a laser

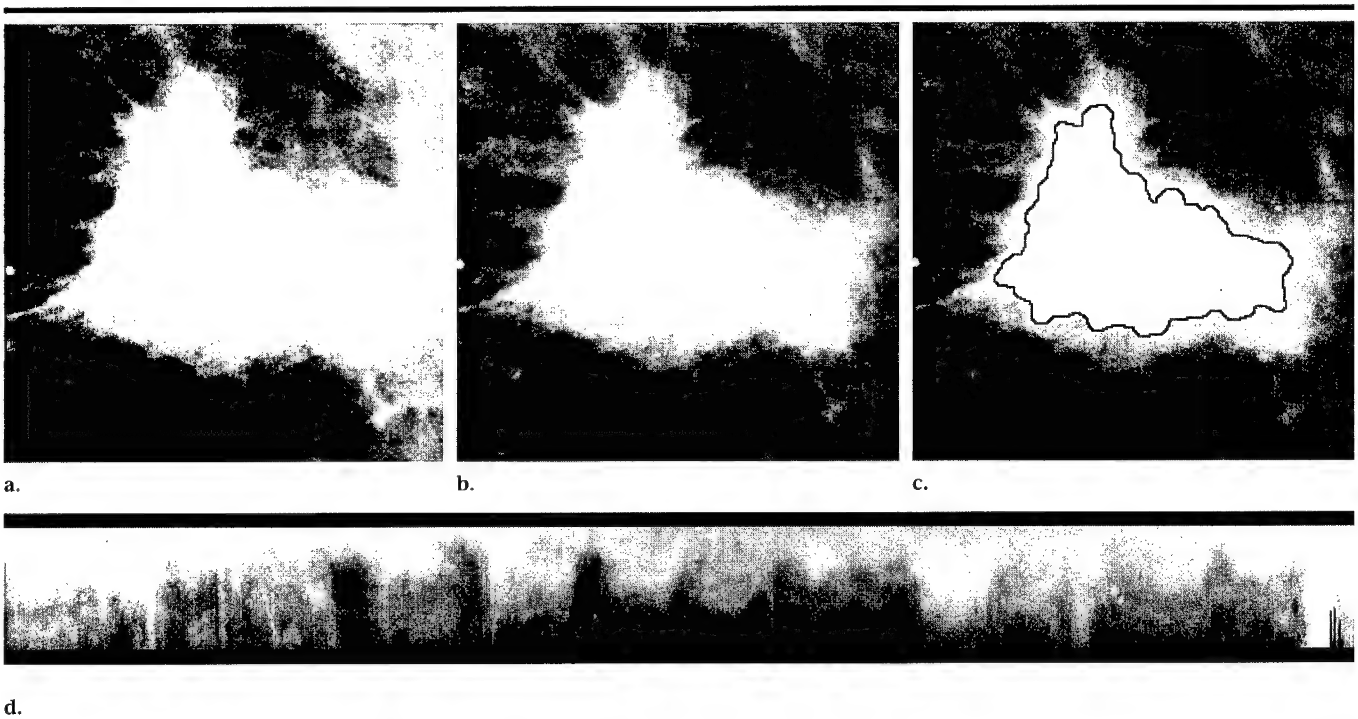


Figure 2. Example of rubber-band-straightening transform for extraction of texture features in the margin region surrounding a mass. **(a)** Original and **(b)** background-corrected images showing the region of interest with the mass, **(c)** mammogram showing an outline of the segmented mass, and **(d)** rubber-band-straightening–transformed image of a 40-pixel-wide region surrounding the segmented mass.

imager (Lumisys DIS-1000, Los Altos, Calif) at a pixel size of 0.1 × 0.1 mm and 12-bit gray levels. This imager has an optical density range of about 0.0–3.5. The optical density on the film was digitized linearly to pixel value at a calibration of 0.001 optical density unit/pixel value in the optical density range of about 0.0–2.8. The digitizer deviated from a linear response at an optical density higher than 2.8.

For the observer experiments, we used laser-printed images of the digitized mammograms for all readings. The images were printed with a 969HQ laser imager (Imation, Oakdale, Minn) that was connected to a Macintosh computer (Apple Computer, Cupertino, Calif) through a special digital interface. The interface provided a 12-bit in, 10-bit out look-up table and allowed images to be scaled to different factors with 15 interpolation methods. Because this laser imager has a pixel size of about 0.085 mm, we enlarged the images by about 18% during printing to maintain them at the same size as the original mammograms. One of the interpolation methods was chosen by an experienced radiologist (M.A.H.), who inspected the printed images with a magnifier and evaluated the sharpness of the spicules and mass boundaries. Because of the small pixel size used for both digitization and printing, basically no noticeable blurring of the masses could be seen with the chosen interpolation method. The images were also inspected for the potential contouring effect of 10-bit output images, but no noticeable artifacts could be found. A linear pixel value–to–output optical density calibration curve of the laser imager was used for the printing. All images were printed with the same settings.

Computerized Classification of Masses

Our computerized method of classifying mammographic masses has been described in detail previously (15–17). The method is summarized as follows: A region of interest that contained the biopsy-proved mass was identified on the mammogram by the radiologist. Background correction based on a distance-weighted estimation method was applied to the region of interest to reduce the low-frequency density variation in the region. A median-filtered smoothed image and two high-frequency enhanced images were generated from the background-corrected region of interest. The smoothed and enhanced gray-level values at each pixel were used as features in a k-means clustering algorithm to classify the pixels into two clusters; one was the mass, and the other was the surrounding breast tissue background. By choosing an appropriate criterion, a mass region slightly smaller than the actual mass that was visible on the image was segmented.

The boundary of the segmented region was smoothed by morphologic filtering. A new image transformation technique, referred to as the rubber-band-straightening transform, was used to transform a 40-pixel-wide region that surrounded the segmented mass boundary into a rectangular region. After transformation, the mass margin became approximately parallel, and any spicules that were radiating from the mass became approximately perpendicular, to the long dimension of the rectangular region. The rubber-band-straightening transform enabled the spicules to be aligned approximately in a uniform direction and thus facilitated the extraction of texture features from the margin of the mass. An example of a rubber-band-straightening–transformed image is shown in Figure 2.

Two types of texture features were found to be useful for classification. The first set of features included eight texture measures derived from the spatial gray-level dependence matrices of the rubber-band-straightening–transformed image. A spatial gray-level dependence matrix ele-

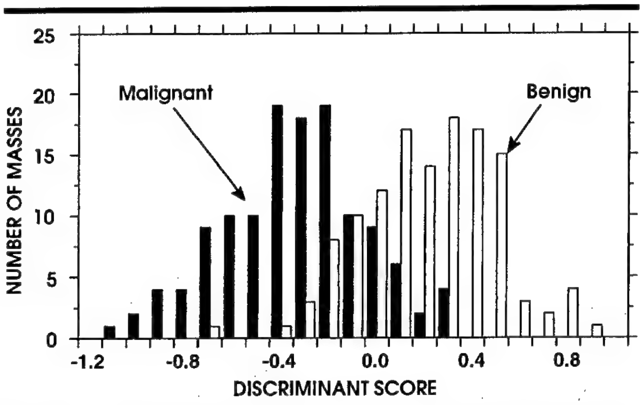


Figure 3. Histogram of the test discriminant scores of the 253 masses obtained from the linear discriminant classifier by using a "leave one case out" training and test resampling scheme. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. The discriminant scores were used as the decision variable in the ROC analysis of classification performance.

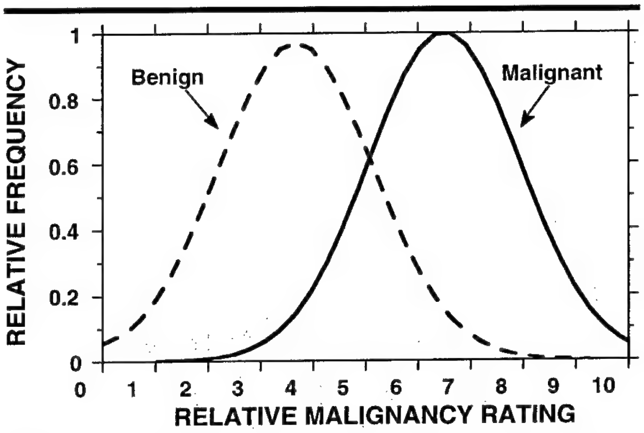


Figure 4. Binormal distribution fitted to the histogram of the discriminant scores of the malignant and benign masses. The discriminant scores were linearly transformed into a relative malignancy rating ranging from 1 to 10, where 1 corresponded to the most benign rating and 10 corresponded to the most malignant rating. This binormal distribution was shown to the observers during the training session to explain the rating scale of the computer classifier.

ment $p_{\theta,d}(i,j)$ is the joint probability of the occurrence of gray levels i and j for pixel pairs that are separated by a distance d and at a direction θ (22). For analysis of the masses, the spatial gray-level dependence matrices were constructed for 10 pixel distances (d = 1, 2, 3, 4, 6, 8, 10, 12, 16, 20 pixels) and in four directions (0°, 45°, 90°, 135°) relative to the mass boundary. Therefore, a total of 320 spatial gray-level dependence texture features were extracted.

The second set of texture features was derived from the run length statistics matrices of the horizontal and vertical gradient images of the rubber-band-straightening–transformed margin region. Five texture measures were extracted from the run length statistics matrix in each of the two directions (0° or 90°) on each gradient image. A total of 20 run length statistics texture features were thus obtained. Therefore, we had a total of 340 features from the two types of texture measures.

A stepwise linear discriminant feature selection procedure (23) was used to select the most effective features from the available feature set. A total of 41 features were selected. The selected features were input into the Fischer linear discriminant classifier (24) as predictor variables. A "leave one case out" resampling scheme was used to train and test the classifier. A histogram illustrating the test discriminant scores of the 253 masses is shown in Figure 3. For this classifier, a smaller discriminant score corresponded to a higher likelihood of malignancy. By using the test discriminant score as the decision variable, the performance of the computer classifier could be evaluated by using ROC analysis (17,25,26) and compared with that of the radiologists, as described later.

Relative Malignancy Rating of the Masses

For the observer performance study, we provided a relative malignancy rating of each mass to the observer during the reading session with CAD. The relative malignancy rating was obtained by taking a linear transformation of the computer classifier's decision variable to a range of 1–10 and rounding the value to the nearest integer. The transformation also reversed the relative magnitude of the decision variables so that 1 corresponded to the highest benignity rating, and 10 corresponded to the highest malignancy rating.

The purpose of the transformation was to provide a simple and intuitive relative scale for the observer. Because the transformation was linear and monotonic, the distributions of the normal and abnormal samples, as well as their ROC curves, were not affected, with the exception of a small error caused by making the decision variables discrete. Furthermore, the slope a and intercept b parameters that were fitted to the transformed discriminant scores for the normal and abnormal samples by using the LABROC program (26) were used to generate a binormal distribution. The fitted binormal distribution with the relative malignancy rating on a 1–10 scale (Fig 4), together with the computer's ROC curve, were shown and explained to the observers during a training session.

Observer Performance Study

Two ROC experiments (27) were conducted: The masses were evaluated from a single view in the first experiment and from two views in the second experiment. The location of the biopsy-proved mass was marked on each image so that the correct mass was evaluated by all observers. The observers were instructed to ignore any other possible masses on the images. Six radiologists (M.A.H., M.A.R., T.E.W., D.D.A., C.P., J.S.N.) who are approved by the Mammography Quality Standards Act and have 7–20 years of experience in interpreting mammograms participated in the observer performance experiments.

There were two reading sessions in each experiment—one with CAD and the other without CAD. The observers were asked to rate the likelihood of malignancy of the masses on a 10-point confidence rating scale under all reading conditions. In the first session, half the observers interpreted the images without CAD, and the other half interpreted them with CAD. The two reading sessions in the same experiment were separated by at least 3 weeks, and the two experiments were separated by 6 months. For all four reading sessions, the observer had unlimited time to read each case. To estimate the average reading time per case for each observer, the reading time for each case was recorded by using a stopwatch.

In the first experiment, the data set of 253 single-view mammograms was divided into a training set of 15 mammograms and a study set of 238 mammo-

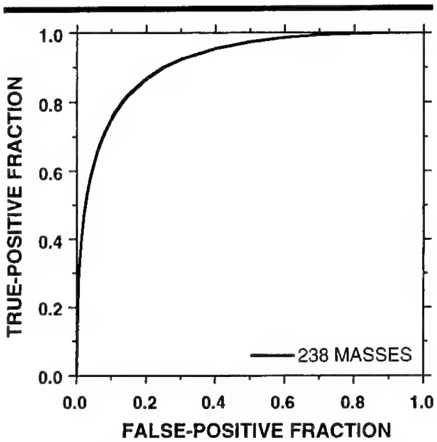


Figure 5. ROC curve for computerized classification of the 238 masses used in the observer performance study with single-view reading. The computer's ROC curve can be compared with the radiologists' ROC curves obtained from the single-view reading experiment illustrated in Figures 6 and 8.

grams (117 benign, 121 malignant). In each reading session, training was conducted before the reading of the study images. For the reading session with CAD, the fitted binormal distributions of the computer rating scores (Fig 4) for the entire data set were explained to the observer during training to familiarize the observer with the computer's rating scale. The computer rating of the mass was displayed on each image. After reading each training image, the observer was told the results of biopsy of the mass.

Each observer read the entire data set in one reading session. The order of the study images was randomized by a random number generator. The random sequence was different for each observer and for each reading session by the same observer. For the reading session with CAD, the observer was free to look at the computer rating, which was displayed on the image, either before or after estimating the likelihood of malignancy of the mass. However, each observer was asked to always read the computer rating before making a final decision. The observer was not informed of the pathologic results of any mass on the study images.

The second experiment was very similar to the first experiment. From the 238 single-view mammograms, 76 matched pairs (37 benign, 39 malignant) of craniocaudal and mediolateral oblique or lateral views were found. Another six pairs of two-view mammograms were identified from the rest of the images and used as training cases. The remaining mammograms were either single-view images or additional views of the pairs already chosen, so they were not used in this experiment. In this experiment, the observers were not informed of the pathologic results of any study case in any reading session. The 76 pairs of mammograms were read in one reading session by each observer.

For the reading session with CAD, the rating of the mass in each view was displayed on the respective image. The computer ratings of the mass on the two views were generally different. It was up to the observer to decide how to merge the two-view information. Observers were asked to give a single rating of the mass after reading both views.

ROC Analysis

The confidence ratings of each observer obtained from each reading condition were analyzed by using ROC methodology, and the classification accuracy was quantified by using the area under the ROC curve, A_z. A maximum likelihood estimation of the binormal distribution was fitted to the confidence ratings by using the LABROC program. This program provides an estimate of the A_z and of the a and b parameters of the ROC curve. The statistical significance of the difference in A_z between the reading with CAD and that without CAD was estimated with two methods: One was the Student paired t test for observer-specific paired data; the other was the Dorfman-Berbaum-Metz method for analysis of multireader, multicase ROC data (28). The statistical significance of the difference in $A_z$for reading single-view and two-view mammograms was estimated by using the Student paired t test for the six observers. The Student paired t test takes into account the statistical variation of readers, whereas the Dorfman-Berbaum-Metz method considers both reader variation and case sample variation by means of an analysis of variance approach. Therefore, the results of Dorfman-Berbaum-Metz analysis can be generalized to the population of readers as well as to the population of case samples.

Positive Predictive Value

An ROC curve represents the entire range of operating conditions of a diagnostic process and is independent of disease prevalence. When the disease prevalence is known, any operating point on an ROC curve can be used to derive the PPV and the corresponding false-negative fraction (false-negative fraction = 1 − true-positive fraction) on the basis of the following relationship: PPV = TPF × P(M)/[TPF × P(M) + FPF × P(B)], where TPF is the true-positive fraction, FPF is the false-positive fraction at the chosen decision threshold, and P(M) and P(B) are the prevalences of malignant and benign cases, respectively. By varying the decision threshold, the dependence of the PPV on the false-negative fraction can be derived.

Because our data set did not include masses on which biopsy had not been performed, the ROC curves obtained in this study cannot be generalized to predict the performance of the computer classifier and radiologists in clinical practice. However, to demonstrate the possible effect of CAD on the PPV in the population of masses in which biopsy is likely to be performed under the current clinical criteria, we can estimate the PPV by using the prevalence of the malignant and benign masses in this patient group. Because the PPV of masses sent for biopsy ranges from about 25% to 44% in general and from about 25% to 30% at our institution, for the purposes of our estimation, we assumed that the P(M) was 25% and the P(B) was 75% in this population. A higher prevalence of malignant cases would cause an increase in the PPV, but the trend between the PPV curves with and without CAD would be similar.

RESULTS

The ROC curve illustrating the performance of the computer classifier for the 238 study mammograms is shown in Figure 5. The ROC curve for the entire set of 253 mammograms (not shown) was almost identical to that of the 238 study cases; this indicates that the 15 training cases were typical of the 238 cases used in the study. The A_z values (± SD) for both ROC curves were 0.92 ± 0.02.

For the first experiment of reading the 238 single-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 6a and 6b, respectively. The A_z values of the six radiologists for the readings with and without CAD are listed in Table 1.

For the second experiment of reading the 76 pairs of two-view mammograms, the ROC curves for the readings by the six radiologists both without and with CAD are shown in Figures 7a and Figure 7b, respectively. The A_z values of the six radiologists in this experiment are also listed in Table 1.

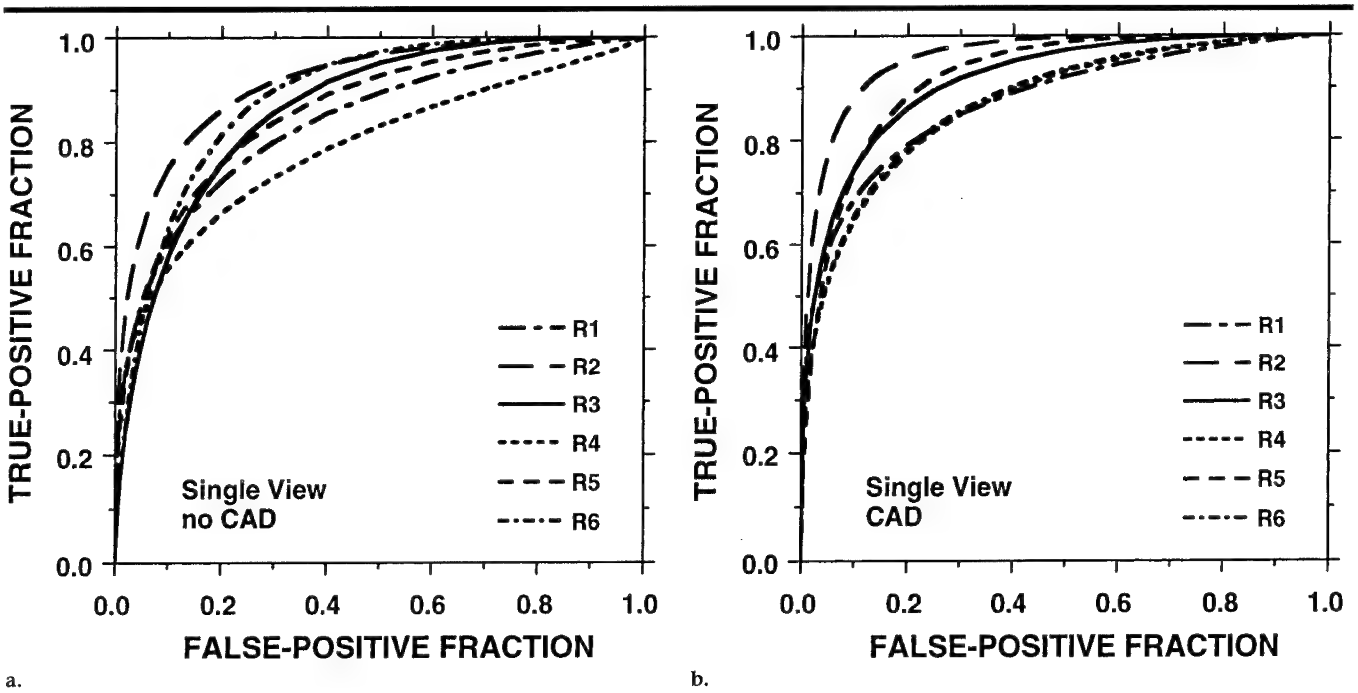


Figure 6. ROC curves for the six observers for single-view reading of the masses **(a)** without CAD and **(b)** with CAD. **(a, b)** *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. Five of the six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

The average ROC curve was derived from the average *a* and *b* parameters of the six individual ROC curves for a given reading condition (27). The average ROC curves for the four reading conditions are shown in Figure 8. The A_z values of the average ROC curves are listed in Table 1.

TABLE 1
Areas under the ROC Curves for the Classification of Masses with and without CAD by the Six Radiologists

Radiologist No.	A_z (Single View)* Without CAD	A_z (Single View)* With CAD	A_z (Two View)† Without CAD	A_z (Two View)† With CAD
1	0.84 ± 0.03	0.87 ± 0.02	0.90 ± 0.03	0.93 ± 0.03
2	0.92 ± 0.02	0.96 ± 0.01	0.95 ± 0.02	0.97 ± 0.02
3	0.86 ± 0.02	0.91 ± 0.02	0.92 ± 0.03	0.93 ± 0.03
4	0.79 ± 0.03	0.87 ± 0.02	0.88 ± 0.04	0.95 ± 0.03
5	0.86 ± 0.02	0.92 ± 0.02	0.93 ± 0.03	0.97 ± 0.02
6	0.89 ± 0.02	0.87 ± 0.02	0.89 ± 0.04	0.93 ± 0.03
A_z from average a, b parameters	0.87	0.91	0.92	0.96

Note.—Data are the mean ± SD.

* P = .022 for the difference between the A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. P = .020 for this difference, as determined by using the Dorfman-Berbaum-Metz method.

† P = .007 for the difference between A_z values measured with CAD and those measured without CAD, as determined by using the Student two-tailed *t* test. P = .026 for this difference, as determined by using the Dorfman-Berbaum-Metz method.

For the reading of the single-view mammograms, the performance of the computer classifier was comparable to that of the radiologist (reader 2) who had the highest classification accuracy (compare Figs 5 and 6) and higher than the average performance of the six radiologists (compare Figs 5 and 8). When the radiologists read the images with the computer aid, the classification accuracy of five radiologists improved (Table 1); the improvement in their A_z values ranged from 0.04 to 0.08. The average performance of the six radiologists became comparable to that of the computer classifier. The improvement in the radiologists' classification accuracy by using CAD was statistically significant (P = .022, Student paired *t* test; P = .020, Dorfman-Berbaum-Metz method). Reader 2 with CAD obtained an A_z value of 0.96, which was higher than that obtained by the radiologist alone or by the computer alone.

A trend similar to that with the single-view readings was observed with the two-view readings. The A_z value of the computer classifier for the corresponding 152 single-view masses was 0.91 ± 0.02. The classification accuracy of all six radiologists improved when they read the mammograms with the computer aid. The increase in the A_z values ranged from 0.01 to 0.07. The improvement was statistically significant (P = .007, Student paired *t* test; P = .026, Dorfman-Berbaum-Metz method). With CAD, two radiologists achieved an A_z value of 0.97, which was higher than that obtained by the radiologists alone or by the computer alone. These results indicate that the second opinion provided by the computer classifier might have strengthened the radiologists' confidence in the interpretation of some difficult cases but had less influence on the radiologists' decision when the computer made mistakes or when the radiologists were confident about their decision.

As can be seen from the data in Table 1,

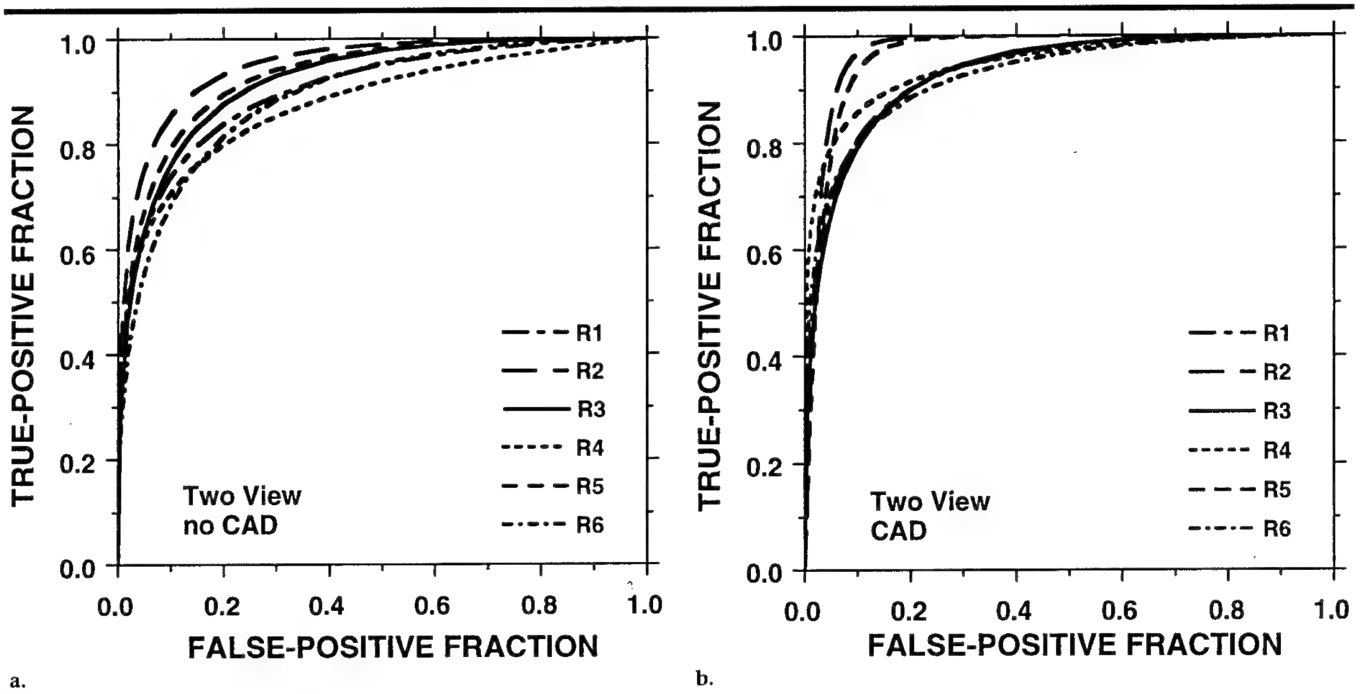


Figure 7. ROC curves for the six observers for two-view reading of the masses (a) without CAD and (b) with CAD. (a, b) *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6. All six observers achieved an increase in the area under the ROC curve, A_z, with CAD.

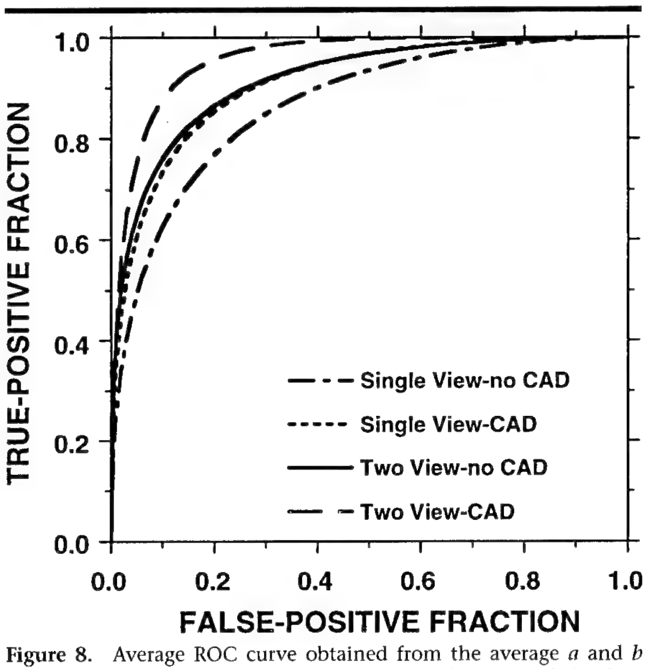


Figure 8. Average ROC curve obtained from the average *a* and *b* parameters of the six individual ROC curves for each of the four reading conditions. An improved ROC curve was achieved with CAD in both the single-view and two-view reading experiments.

the radiologists' accuracy in classifying masses by reading two-view mammograms was consistently higher than that by reading single-view mammograms (P = .008). This trend remained when they read the mammograms with CAD (P = .007). These findings are consistent with the clinical experience of the radiologists that at least two views of mammograms are needed to effectively evaluate a suspicious lesion.

The PPV as a function of the false-negative fraction was derived from the fitted ROC curves under the assumption that the prevalence of malignant masses was 25% in the population of masses sent for biopsy. The PPVs estimated for the six observers who read the two-view mammograms with and without CAD are plotted in Figure 9. CAD would provide an improvement in the PPV in the high false-negative fraction range for all observers except readers 2 and 5. The increase in the PPV at a decision threshold of "no missed malignant mass" (ie, false-negative fraction = 0) varied over a wide range; the largest gain, 39%, would be achieved by reader 2, and the smallest gain, 0%, would be achieved by reader 4.

DISCUSSION

In the observer experiment of reading two-view mammograms with CAD, we presented the computer's rating of each view separately. The decision of how to merge the computer ratings of the two views was left to the radiologist. It is likely that the radiologists took the conservative approach of using the highest malignancy rating of the two as the computer's overall rating. However, it also might have depended on whether the relative ranking between the two computer ratings agreed with the observer's opinion. In some cases, we observed that the radiologist's rating was very different from the computer's rating of either view.

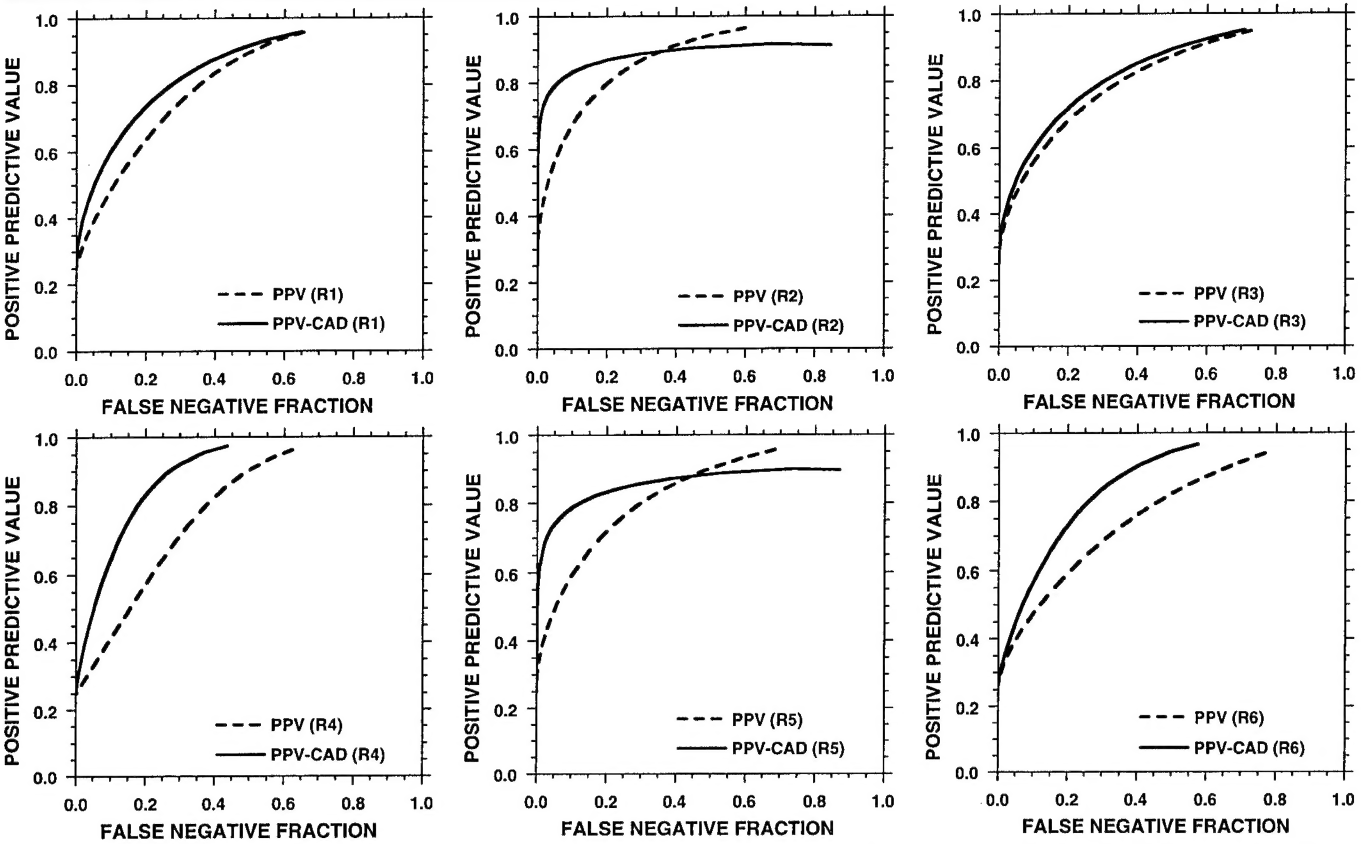


Figure 9. PPV as a function of the false-negative fraction derived from the ROC curves for the six observers (Fig 7). The PPV was predicted for a population of masses in which biopsy was likely to be performed under current clinical criteria and by assuming the prevalence of malignant masses to be 25%. *R1* = reader 1, *R2* = reader 2, *R3* = reader 3, *R4* = reader 4, *R5* = reader 5, *R6* = reader 6.

Because decision making is a complex process, the simple approach of using the highest malignant rating or the average rating from multiple views may not be the method preferred by radiologists. The separate ratings that we used in this study would provide less biased information. Further investigation is needed to determine the best approach of presenting the computer's ratings to radiologists in clinical practice.

To obtain insight into how the radiologists might use the two-view information, we compared the classification results from their true two-view reading with those from a simulated two-view reading without the computer aid. The latter results were derived from ratings of single-view readings of the same 76 pairs of mammograms interpreted in experiment 2 by assuming two strategies—one in which the highest malignancy rating between the two ratings was used, and the other in which the average of the two ratings was used (Table 2). The A_z values for these classification ratings derived from the single-view reading are listed in Table 2. The corresponding A_z values for the computer classifier are also given in Table 2 for comparison.

The A_z values for the maximal rating and the average rating were similar. Four of the radiologists obtained higher A_z values at the true two-view reading; the A_z values obtained by the remaining two radiologists were lower than those obtained at the simulated two-view reading. Although the difference did not achieve statistical significance ($P = .37$) and both readings included intraobserver variations, there seemed to be a slight trend toward the true two-view reading being more accurate than the simulated two-view reading. This may indicate that the radiologists used a more complex decision-making process to interpret the two views of the masses than that of simply maximizing or averaging the ratings from each view.

In this study, the discriminant scores of the masses given by the computer classifier were transformed into a relative malignancy rating. The relative malignancy rating scale and the distribution of the malignant and benign masses along the relative rating scale were explained to the observers in the training sessions. A relative malignancy rating scale was used because the true likelihood of malignancy of the masses could not be estimated from a small data set, as will be explained. However, the relative rating scale provided by the computer was ad-

TABLE 2
Estimation of the Malignancy Classification of 76 Masses by Two-View Reading, as Simulated from Single-View Reading of Mammograms by Radiologists without CAD

	A_z	
Radiologist No.	Maximal Rating	Average Rating
1	0.94 ± 0.03	0.93 ± 0.03
2	0.94 ± 0.03	0.94 ± 0.03
3	0.84 ± 0.05	0.86 ± 0.04
4	0.85 ± 0.04	0.83 ± 0.05
5	0.88 ± 0.04	0.89 ± 0.04
6	0.91 ± 0.03	0.92 ± 0.03
Computer	0.96 ± 0.02	0.96 ± 0.02

Note.—Data are the mean ± SD. Two strategies were used: In one, the highest of the malignancy ratings on each view was used; in the other, the average between the two ratings was used.

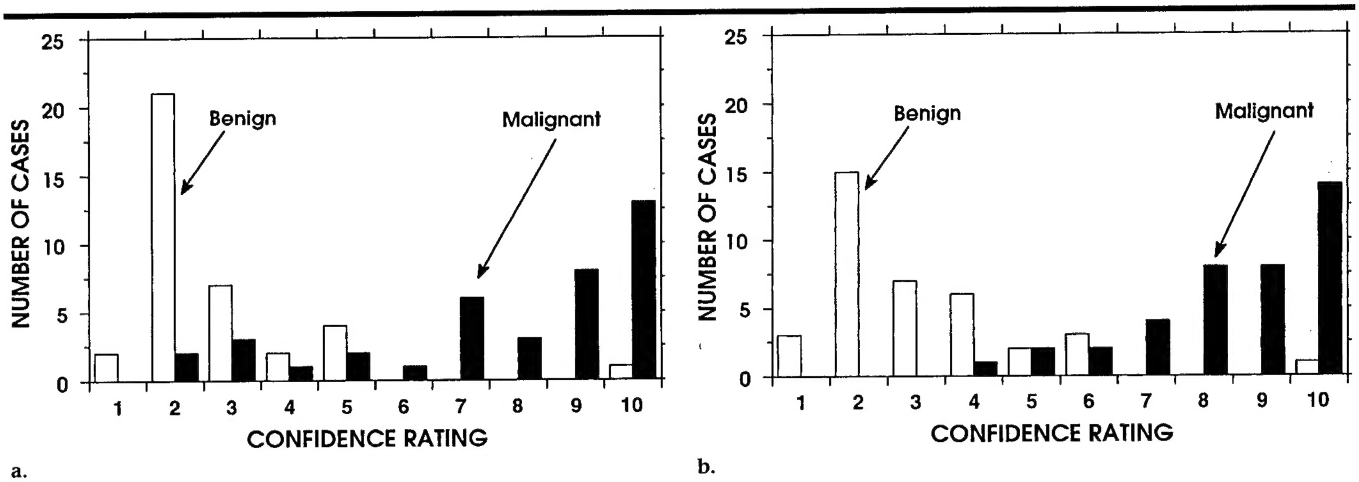


Figure 10. Histograms illustrate the confidence ratings of reader 5 obtained by reading 76 two-view mammograms (a) without CAD and (b) with CAD. The specificity of reader 5 at 100% sensitivity would increase from 5% (two of 37 masses) without CAD to 68% (25 of 37 masses) with CAD if an appropriate decision threshold were chosen.

equate for measuring the relative performance of classification with and without CAD in an ROC study.

If a computer classifier is trained and tested with very large data sets, and if both the malignant and benign cases represent random samples of the population, then the likelihood of malignancy of a classified mass can be estimated on the basis of the probability distributions of the classifier's test output scores and the prevalence of the two classes of masses in the patient population. However, with a relatively small data set, such as that used in this and other observer studies (14), there are limitations. First, the performance of a classifier trained with a small sample set may have large bias and variance (29–31). Second, the data set in this study did not include masses on which biopsy was not performed, so it did not represent a random sample of the masses in the patient population. If our classifier were applied to all cases of solid masses in clinical practice, the probability distribution of the test scores for the two classes of masses would be different from that of the current data set.

If we ignore the patient population at large, it is possible to estimate the likelihood of malignancy of a mass on the basis of the probability distribution of the classifier output scores by using the prevalence of the two classes of masses in this specific data set. However, the likelihood of malignancy derived in this way will be completely different from the true likelihood of malignancy of a mass in the patient population. This can be easily seen if one considers that the same mass with the same discriminant score will have a smaller likelihood of malignancy if it is analyzed within a data set that has a lower prevalence of malignant cases than that in the current data set.

Training the participating radiologists with a "likelihood of malignancy" derived from a small data set for the observer experiment may mislead them if they encounter a similar mass in their clinical practice. We, therefore, preferred to use a "relative malignancy rating," which is independent of the prevalences of malignant and benign masses in the data set. As long as the same classifier and the same linear transformation are used for classifying masses, the relative malignancy rating for a given mass will remain the same, regardless of the types of other masses in the data set. When a computer classifier is implemented in a clinical setting and its performance can be established in the patient population, the true likelihood of malignancy of a given mass can be estimated and provided to the radiologist. The true likelihood of malignancy may be a more informative measure for radiologists in the clinical application of CAD.

For the reading of the 76 two-view mammograms, the results of the ROC study indicated an improvement in the A_z value for all six radiologists when the computer aid was used. This indicates an overall increase in the separation of confidence rating distributions between the malignant and benign cases. The histograms in Figure 10 illustrate the distributions of confidence ratings with and without CAD for reader 5, who achieved the second greatest improvement in both the A_z value (Table 1) and the separation of malignant from benign distributions. Without CAD, this reader's ratings of the malignant cases ranged from 2 to 10. This is consistent with the fact that biopsy was performed in all masses in the data set to avoid missing the malignant cases. With CAD, there was marked improvement in the separation of the two distributions. It is possible to set a decision threshold at a confidence rating of 4, below which biopsy would not need to be performed and no malignant masses would be missed. The number of benign masses that could be identified without missing a malignant mass by setting an appropriate threshold would increase by 23 (out of 76 cases) for reader 5. Five of the six radiologists in our ROC study achieved an improvement in distinguishing benign from malignant masses, and one radiologist had no difference. Although the improvement of the five radiologists varied over a wide range, from one to 25 cases, this result indicates a strong possibility that CAD can be used to reduce the number of unnecessary biopsies.

The large variation in improvement among the radiologists may have been due to the different degrees of confidence that they had in the computer aid. As with any new diagnostic tool, this confidence is influenced by the experience the radiologist has with the tool. Although the radiologists received training before the reading sessions, the high variability in confidence was not unexpected, because this ROC study was the first instance in which they had worked with the computer aid. Their confidence levels may have also been reflected in the relatively low accuracy of classification by some radiologists with CAD compared with that of the computer classifier alone.

If a radiologist can increase his or her confidence in the performance of a computer aid by gaining more extensive clinical experience, then he or she will likely be able to find the most effective way of merging his or her judgment with the computer's rating and thus reduce both interobserver and intraobserver variability. Because a radiologist who uses CAD can establish a meaningful decision threshold for biopsy only after becoming familiar with the sensitivity and specificity of working with CAD, the radiologists in this study were not asked to decide whether biopsy should have been performed on a mass. Rather, we focused on the evaluation of changes in the sensitivity and specificity of the radiologists' classification of masses when CAD was used.

In this ROC study, all six observers were attending radiologists with extensive experience in the interpretation of mammograms. It is possible that the computer aid may be even more useful to radiology residents or radiologists with less experience in mammography. The effect of CAD on mammographic interpretation by less-experienced readers will be a subject of investigation in future studies.

The observers were allowed unlimited time to read each case in this ROC study. To obtain an estimate of the change in reading time with CAD, we recorded the reading time of each observer in each reading session by using a stopwatch. For the single-view reading experiment, the average reading time per image without CAD varied from 4.3 seconds to 17.1 seconds (mean time for the six observers, 7.8 seconds). The average reading time per image with CAD varied from 4.2 seconds to 17.3 seconds (mean time, 7.3 seconds). For the two-view reading experiment, the average reading time per pair of images without CAD varied from 6.6 seconds to 16.0 seconds (mean time, 10.4 seconds). The average reading time per pair of images with CAD varied from 7.6 seconds to 27.1 seconds (mean time, 13.5 seconds).

The reading time essentially did not change with use of the computer aid for the single-view readings. For the two-view readings, the radiologists took longer with CAD, probably because they had to merge the two computer ratings and merge the computer ratings with their own evaluations. Further investigation is needed to determine whether there is a trade-off between the radiologist's efficiency and the method of presenting the computer rating and whether the reading time with CAD will depend on the experience that the radiologist has with the computer information.

In the observer study, we used laser-printed mammograms instead of the original mammograms for the reading experiments. A major reason is that it is difficult to keep all the original mammograms together for the entire period of the study because they are part of active patient files and thus often recalled for comparison with new studies or for other clinical reasons. Because the maximum optical density of laser-printed images was 3.1 for the laser imager used, the contrast on the printed mammograms was about 20% lower than that on the original mammograms. Although the image quality was slightly lower than that of the original, the laser-printed digitized images were judged to be adequate for reading the details of the masses by the participating radiologists. The laser-printed image set might also be considered as one that had slightly more subtle masses than the original set of images. Because the relative performance of two modalities is measured in ROC experiments, and because the readings both with and without CAD in this study were conducted with the same set of printed images, the relative performance of the two readings should be valid. It should also be noted that in order for a computer aid that uses automated image analysis to be widely accepted, direct digital mammography would have to be the imaging modality in clinical use. Laser-printed images or soft-copy monitors will be the display medium for the digital mammograms. The use of laser-printed images for this ROC study was therefore practical.

In our observer performance experiment, we found that CAD improved the radiologists' ability to distinguish malignant and benign masses. This is consistent with the results of other studies (11,14) in which a statistically significant improvement ($P < .001$ in both studies) in the radiologists' classification accuracy by using CAD was found. The results of the former study (11) further showed that the PPV of a recommendation for biopsy by the radiologists was significantly increased ($P < .001$). In our approach, the computer classifier automatically extracted image features, whereas in the other studies, the computer classifier used the radiologist's evaluation and other patient information as input. Therefore, it appears that CAD can provide a useful second opinion to radiologists, either by consistently extracting and analyzing the image features or by optimally weighting various diagnostic factors and thereby improving the consistency in the decision-making process. This suggests that a computer classifier that combines both approaches—that is, automatically extracts image features and optimally merges them with the radiologist's evaluation and patient information—may be even more effective for breast cancer diagnosis. The latter step will also improve the radiologist's utilization of the computer rating on the basis of the computer-extracted features; this utilization was found to have large interobserver variation in our ROC experiment.

In conclusion, an ROC study of the effects of CAD on radiologists' classification of malignant and benign masses on mammograms was conducted. The results showed that CAD can provide a statistically significant improvement in the classification accuracy—that is, in the A_z value—for both single-view reading (P = .022) and two-view reading (P = .007). The improved separation between the confidence ratings of the malignant masses and those of the benign masses indicates the potential that CAD may reduce the rate of biopsy of benign masses when decision thresholds are properly chosen by the radiologists. The decision threshold may vary among radiologists, as in the case of mammographic interpretation without CAD, and can be set after the radiologist working with CAD has established his or her sensitivity and specificity with this approach through clinical experience.

Further studies are needed to evaluate the effects of CAD on the accuracy of radiologist classification of masses in clinical settings in which the prevalence of malignant masses is different from that in a laboratory data set and the likelihood of malignancy of a mass can be estimated by the computer classifier. In the two-view reading ROC experiment, the reading time per case increased by about 30% with the use of CAD. The dependence of the radiologist's efficiency in reading with CAD on the presentation method and on the reader's experience in using the computer information also warrants further investigation.

Acknowledgments: The authors are grateful to Charles E. Metz, PhD for useful discussions and for the use of the LABROC and LABMRMC programs.

References

1. Landis SH, Murray T, Bolden S, Wingo PA. Cancer statistics 1998. CA Cancer J Clin 1998; 48:6–29.
2. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr Opin Radiol 1992; 4:123–129.

3. Kopans DB. The positive predictive value of mammography. AJR 1991; 158:521–526.
4. Shtern F. Digital mammography and related technologies: a perspective from the National Cancer Institute. Radiology 1992; 183: 629–630.
5. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
6. Vyborny CJ. Can computers help radiologists read mammograms? Radiology 1994; 191:315–317.
7. Chan HP, Doi K, Vyborny CJ, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990; 25:1102–1110.
8. Kegelmeyer WP, Pruneda JM, Bourland PD, Hillis A, Riggs MW, Nipper ML. Computer-aided mammographic screening for spiculated lesions. Radiology 1994; 191: 331–337.
9. Getty DJ, Pickett RM, D'Orsi CJ, Swets JA. Enhanced interpretation of diagnostic images. Invest Radiol 1988; 23:240–252.
10. D'Orsi CJ, Getty DJ, Swets JA, Pickett RM, Seltzer SE, McNeil BJ. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. Radiology 1992; 184:619–622.
11. Baker JA, Kornguth PJ, Lo JY, Floyd CE. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 1996; 198:131–135.
12. Chan HP, Sahiner B, Petrick N, et al. Observer performance study of radiologists' reading of mammographic masses and comparison with computerized classification (abstr). Radiology 1996; 201(P):370.
13. Chan HP, Sahiner B, Helvie MA, et al. Effects of computer-aided diagnosis (CAD) on radiologists' classification of malignant and benign masses on mammograms: an ROC study (abstr). Radiology 1997; 205(P):275.
14. Jiang Y, Nishikawa R, Schmidt RA, Metz CE, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis (CAD): an observer study (abstr). Radiology 1997; 205(P):274.
15. Sahiner B, Chan HP, Petrick N, Helvie MA, Adler DD, Goodsitt MM. Classification of masses on mammograms using rubber-band straightening transform and feature analysis. Proc SPIE 1996; 2710:44–50.
16. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: the rubber-band straightening transform and texture analysis. Med Phys 1998; 25:516–526.
17. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. Phys Med Biol 1998; 43:2853–2871.
18. Ackerman LV, Gose EE. Breast lesion classification by computer and xeroradiograph. Cancer 1972; 30:1025–1035.
19. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. IEEE Trans Med Imaging 1993; 12:664–669.
20. Pohlman S, Powell KA, Obuchowski NA, Chilote WA, Grundfest-Broniatowski S. Quantitative classification of breast tumors in digitized mammograms. Med Phys 1996; 23:1337–1345.
21. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 1998; 5:155–168.
22. Haralick RM, Shanmugam K, Dinstein I. Texture features for image classification. IEEE Trans Syst Man Cybernetics 1973; 3:610–621.
23. Norusis MJ. SPSS for Windows release 6: professional statistics. Chicago, Ill: Statistical Product for Service Solutions, 1993.
24. Lachenbruch PA. Discriminant analysis. New York, NY: Hafner, 1975; 8–19.
25. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
26. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Stat Med 1998; 17:1033–1053.
27. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
28. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992; 27:723–731.
29. Fukunaga K, Hayes RR. Effects of sample size on classifier design. IEEE Trans Pattern Analysis and Machine Intelligence 1989; 11:873–885.
30. Chan HP, Sahiner B, Wagner RF, Petrick N, Mossoba J. Effects of sample size on classifier design: quadratic and neural network classifiers. Proc SPIE 1997; 3034:1102–1113.
31. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis in mammography: effects of finite sample size. Med Phys 1997; 24:1034–1035.